An information-theoretic approach to assess practical identifiability of parametric dynamical systems

Sanjay Pant, Damiano Lombardi

An information-theoretic approach to assess practical identifiability of parametric dynamical systems.

Sanjay Pant & Damiano Lombardi

INRIA Paris-Rocquencourt, 78153 Le Chesnay, France
UPMC Université Paris 6, Laboratoire Jacques-Louis Lions, 75005 Paris, France

E-mail: {Sanjay.Pant, Damiano.Lombardi}@inria.fr

29 December 2014

Abstract. A new approach for assessing parameter identifiability of dynamical systems in a Bayesian setting is presented. The concept of Shannon entropy is employed to measure the inherent uncertainty in the parameters. The expected reduction in this uncertainty is seen as the amount of information one expects to gain about the parameters due to the availability of noisy measurements of the dynamical system. Such expected information gain is interpreted in terms of the variance of a hypothetical measurement device that can measure the parameters directly, and is related to practical identifiability of the parameters. If the individual parameters are unidentifiable, correlation between parameter combinations is assessed through conditional mutual information to determine which sets of parameters can be identified together. The information theoretic quantities of entropy and information are evaluated numerically through a combination of Monte Carlo and k-nearest neighbour methods in a non-parametric fashion. Unlike many methods to evaluate identifiability proposed in the literature, the proposed approach takes the measurement-noise into account and is not restricted to any particular noise-structure. Whilst computationally intensive for large dynamical systems, it is easily parallelisable and is non-intrusive as it does not necessitate re-writing of the numerical solvers of the dynamical system. The application of such an approach is presented for a variety of dynamical systems – ranging from systems governed by ordinary differential equations to partial differential equations – and, where possible, validated against results previously published in the literature.

Keywords: identifiability, entropy, mutual information, conditional mutual information, uncertainty quantification, non-parametric estimation, dynamical systems, computational mechanics.

Submitted to: ...

1. Introduction

This work is devoted to the setting up of a semi-empirical framework to assess identifiability in parameter estimation problems.

The identifiability problem was first stated by Bellman in [1] and it is a key issue in inverse problems (see [2]). It is still an open problem from a mathematical point of view, when a generic system described by a Partial Differential Equation (PDE) is investigated. The identifiability question can be summarised as follows: *"given a system described by a set of equations and parametrised by certain scalar quantities and a set of measurements of the system, called observables, is it possible to determine the values of the parameters that account for the observables?"*

For Ordinary Differential Equation (ODE) systems, several approaches are proposed in the literature to study identifiability (see [3] for a comprehensive review). Among them, the methods based on an algebraic approach, such as power expansion ([4]), differential algebra (see for instance [5]) and Gröbner basis (see [6, 7]) proved to be a valuable tool. The analyses based on differential algebra investigate the parameters-to-observable by assuming that the measurements are not affected by noise.

Other approaches, as the one proposed in the present work, are based on the construction of databases of pre-computed solutions or on the application of numerical methods in order to assess pragmatical identifiability (see [8]).

The proposed approach is based on a Bayesian framework. Its application in the context of inverse problems is commented in detail in [9]. Recently, it has also been exploited for model selection problems [10] and for experimental optimal design (see, for instance, [11, 12]). A work on the use of Fisher information to assess parameter identifiability is proposed in [13].

The present work aims at setting up a generic framework that could be useful for Computational Mechanics and Biophysical applications. A database of precomputed simulations is computed. In Uncertainty Quantification (UQ) [14–16], the moments of the model outputs are computed when a structured uncertainty in the parameters is assumed. Instead, in the present work, the parameters-to-solution map is studied by means of entropies. These are estimated by using a Monte Carlo (MC) approach.

The use of entropies have been studied extensively in Information Theory since the pioneering work in [17]. The main goal is to set up a framework that allows to assess the identifiability from a pragmatical point of view, by taking into account the *a priori* knowledge about the parameters and the noise level in the measurements. The resulting approach is quite general and can be applied to a large class of systems. Although the approach can be computationally expensive (depending on the computational complexity of the dynamical system under consideration), it is easily parallelisable.

The structure of the work is as follows. In sections 2, 3, 4, and 5, the definition of the entropies and a general discussions about the key quantities involved is presented. Then, the numerical method, its implementation, and its scalability, are detailed in section 6. In the last part several numerical experiments of increasing complexity are

presented. First, a Windkessel model, a commonly employed boundary condition in haemodynamics simulations, is investigated. Then, a non-linear epidemiological model is studied and the results are compared to those presented in [6]. The last two examples deal with PDE systems. The first example concerns the identification of potential in a system of harmonic waves, and the second example deals with parameter estimation in an Advection-Diffusion system.

2. Quantifying information gain through entropy

In the context of parameter estimation of dynamical systems a reasonable question to ask is the following: *"How much information does one expect to gain about the parameters of interest through noisy measurements of an observable (or a set of observables) at a given set of discrete time instants?"* A prerequisite to answer this question is to establish what is meant by *information* about a parameter. In a Bayesian context one may treat all the parameters as random variables and *a priori*, *i.e.* before any measurements are taken, express the beliefs about these parameters in the form of a *prior* probability distribution. Following the acquisition of measurements, one typically updates such beliefs using Bayes' rule to obtain the *posterior* probability distribution. This posterior reflects updated beliefs about the parameters based on both prior information and the measurements. If one can quantify the uncertainty in both the prior and posterior probability distributions then the difference between them can provide a reasonable quantification of the amount of information gained by the measurements. A precursory requirement, hence, is to quantify the uncertainty of a random variable with a given probability density function.

Shannon [17] introduced a measure of uncertainty or 'missing information' for discrete probability distributions that is widely used in the field of *information theory*. His measure of uncertainty, referred as Shannon entropy, is based on three intuitive notions about uncertainty: continuity, monotonic increase with increasing uncertainty, and the composition law (see [17] for details). For a discrete set of probabilities $P_1, P_2, \ldots, P_n$ Shannon entropy, denoted by H, is defined as

$$S = \sum_{i}^{n} P_i \ \log\left(\frac{1}{P_i}\right). \tag{1}$$

The units of Shannon entropy depend on the base of the logarithm used; for base 2, entropy is measured in *bits*, while if natural base is used then entropy is measured in *nats*. For a continuous random variable X with a probability density function $p_X(x)$, the analogue of Shannon entropy, often referred as *differential entropy* is defined as

$$H^S(X) = H^S\left(p_X(x)\right) = \int_{\mathcal{X}} p_X(x) \ \log\left(\frac{1}{p_X(x)}\right) \ dx. \tag{2}$$

where $\mathcal{X}$ is the support of $p_X(x)$. The continuous version does not have the intuitive and desirable properties of the discrete version. The two notable drawbacks associated

with the notion of uncertainty as defined by equation (2) are: first, that if x is not dimensionless then $p_X(x)$ has dimensions and taking log of a dimensional quantity presents problems; and second, that the uncertainty measure is not invariant under a change of variables. Jaynes [18] showed that a renormalised continuous limit of equation (1) that overcomes the drawbacks of equation (2) is

$$H^J(X) = H^J\left(p_X(x)\right) = \int_{\mathcal{X}} p_X(x)\ \log\left(\frac{m(x)}{p_X(x)}\right)\ dx, \tag{3}$$

where $m(x)$ is an 'invariant measure' function introduced by Jaynes. Kullback and Leibler [19] (see also [20]) proposed that a more fundamental measure of information, referred as the discrimination information [21] or generally the Kullback-Leibler (KL) distance/divergence, between two probability distributions $p_X(x)$ and $q_X(x)$ is

$$D\left(p_X(x)||q_X(x)\right) = \int_{\mathcal{X}} p_X(x)\ \log\left(\frac{p_X(x)}{q_X(x)}\right)\ dx. \tag{4}$$

The above expression represents the mean information, i.e. averaged over all outcomes x from $p_X(x)$, for discriminating between $p_X(x)$ and $q_X(x)$. Similar to the expression by Jaynes, equation (3), the expression by Kullback and Leibler, equation (4), is invariant under transformations of the random variable under consideration. The KL-distance can also be viewed as the loss of information if $q_X(x)$ is used to approximate $p_X(x)$. Alternatively, it can be viewed as the information gained about the random variable when its probability distribution changes from $q_X(x)$ to $p_X(x)$ [22]. Hobson [23] proved that the KL-distance of equation (4) satisfies all the intuitively reasonable properties desired in an information measure including those originally proposed by Shannon in [17]. Hobson and Cheng [22] argued that the uncertainty (missing information) in a probability distribution can be obtained by the KL-distance as follows: if $p_X^m(x)$ denotes the distribution with maximum information content (subject to the constraints of the random variable X), and $p_X^o(x)$ denotes the distribution with minimum information content, then the missing information in a probability distribution $p_X(x)$ can be seen as the difference between the information gain when the probability distribution changes from $p_X^o(x)$ to $p_X^m(x)$ and the information gain when the probability distribution changes from $p_X^o(x)$ to $p_X(x)$

$$H^H(X) = H^H\left(p_X(x)\right) = D(p_X^m(x)||p_X^o(x)) - D(p_X(x)||p_X^o(x)) \tag{5}$$

$$= \int_{\mathcal{X}} p_X^m(x)\ \log\left(\frac{p_X^m(x)}{p_X^o(x)}\right)\ dx - \int_{\mathcal{X}} p_X(x)\ \log\left(\frac{p_X(x)}{p_X^o(x)}\right)\ dx. \tag{6}$$

They also showed that Shannon's measure of equation (2), Jaynes' measure of equation (3), and the measure of equation (5) based on the KL-distance, are identical (except for different scales), when the prior probability distribution $p_X^o(x)$ is uniform and viewed as the invariant measure function $m(x)$ of Jaynes.

While the quantities H^S, H^J, and H^H can be used to quantify the inherent uncertainty of a random variable, the quantity of interest in this article is the gain in information (reduction in uncertainty). Consider a random variable X and a related

observable Y (note that Y, too, is a random variable). Prior to any measurement of Y, let the probability distribution of X be $p_X(x)$, the prior probability distribution. Through a measurement $Y = y$ one can employ Bayes' theorem to update the distribution of X from $p_X(x)$ to $p_{X|Y}(x|y)$, the posterior distribution. Note that this posterior is a distribution of X, and while writing it as $p_{X|Y=y}(x)$ or $p_X(x|Y = y)$ is more informative, it is denoted as $p_{X|Y}(x|y)$ for notational simplicity. Through Bayes' theorem one may write $p_{X|Y}(x|y)$ as

$$p_{X|Y}(x|y) = \frac{\overbrace{p_{Y|X}(y|x)}^{\text{likelihood}} \overbrace{p_X(x)}^{\text{prior}}}{\underbrace{p_Y(y)}_{\text{evidence}}}. \tag{7}$$

Consequently, the gain in information about X through the measurement $Y = y$ can be written as

$$G_{X|Y=y}(y) = H\left(p_X(x)\right) - H\left(p_{X|Y}(x|y)\right) \qquad H \in \{H^S, H^J, H^H\} \tag{8}$$

where $G_{X|Y=y}(y)$ denotes the aforementioned gain in information (measured through corresponding H: H^S, H^J, or H^H) and is a function of the measurement y. For an *a priori* analysis, i.e. without any real measurement of Y available, the average/expected gain in information is considered. This expected gain in information, $I_{X|Y}$, can be obtained by integrating $G_{X|Y=y}(y)$ over all values of Y

$$I_{X|Y} = \int_{\mathcal{Y}} G_{X|Y=y}(y)\ p_Y(y)\ dy, \tag{9}$$

where $\mathcal{Y}$ is the support of Y. Although different measures H (H^S, H^J, H^H) lead to different quantification of the inherent uncertainty in a random variable, all these measures, including the measure of KL-distance, lead to identical quantification of expected gain in information when the probability distribution of X changes from $p_X(x)$ to $p_{X|Y}(x|y)$. Proof of this proposition is presented in appendix Appendix A. Since the continuous form of Shannon entropy, the differential entropy of equation (2), has the simplest form, it is chosen to measure the gain in information. In what follows, the symbol H always denotes this continuous form of Shannon entropy. Hence, $H(X)$ denotes the entropy of the marginal/prior probability distribution of X and is calculated as

$$H(X) = H\left(p_X(x)\right) = \int_{\mathcal{X}} p_X(x)\ \log\left(\frac{1}{p_X(x)}\right)\ dx. \tag{10}$$

Furthermore, $H\left(X|Y = y\right)$ is used to denote $H\left(p_{X|Y}(x|y)\right)$, and $H(X|Y)$ is used to denote the expected value of $H\left(p_{X|Y}(x|y)\right)$ with respect to the random variable Y, i.e. with respect to $p_Y(y)$

$$H(X|Y) = \int_{\mathcal{Y}} H\left(X|Y = y\right)\ p_Y(y)\ dy = \int_{\mathcal{Y}}\int_{\mathcal{X}} p_{X|Y}(x|y)\ \log\left(\frac{1}{p_{X|Y}(x|y)}\right)\ p_Y(y)\ dx\ dy$$

$$H(X|Y) = \int_{\mathcal{Y}}\int_{\mathcal{X}} p_{X,Y}(x,y)\ \log\left(\frac{1}{p_{X|Y}(x|y)}\right)\ dx\ dy, \tag{11}$$

where $p_{X,Y}(x,y)$ denotes the joint distribution of the random variables X and Y. By equations (9), (10), and (11), the expected reduction in uncertainty, $\Delta H(X|Y)$, or equivalently the gain in information, $I_{X|Y}$, about X by observing Y can be written as

$$\Delta H(X|Y) = I_{X|Y} = H(X) - H(X|Y) \tag{12}$$

Remark 2.1. The quantity $\Delta H(X|Y)$ or $I_{X|Y}$ is also known as *mutual information*. The mutual information between two random variables X and Y is a familiar concept in information theory and is interpreted as the average amount of information X carries about Y, and vice versa. In this manuscript it is denoted as $M(X;Y)$. It is usually defined in the following form, which is easy to derive from equations (10), (11), and (12)

$$M(X;Y) = \int_{\mathcal{Y}} \int_{\mathcal{X}} p_{X,Y}(x,y) \ \log\left(\frac{p_{X,Y}(x,y)}{p_X(x)\ p_Y(y)}\right) \ dx\ dy. \tag{13}$$

Hence, the following three quantities are equivalent

$$\Delta H(X|Y) = I_{X|Y} = M(X;Y) \tag{14}$$

Remark 2.2. Mutual information is also the KL-distance, see equation (4), between the joint distribution $p_{X,Y}(x,y)$ and the joint distribution of X and Y if they were independent. In the latter case the joint distribution factorises into the product of the marginal distributions, i.e. $p_{X,Y}(x,y) = p_X(x)\ p_Y(y)$ if X and Y are independent.

Remark 2.3. Mutual information is non-negative, $M(X;Y) \geq 0$ (see [24] for proof).

Remark 2.4. Mutual information is invariant under homeomorphic transformations of X and Y . If $X' = P(X)$ and $Y' = Q(Y)$, where P and Q are smooth and uniquely invertible maps then $M(X;Y) = M(X';Y')$ (see [25] for proof).

Remark 2.5. While the above theory is presented for scalar random variables X and Y, it also extends to vector random variables. For example, the continuous shannon entropy of a vector random variable $\mathbf{X}$ with probability distribution function $p_{\mathbf{X}}(\mathbf{x})$ is $H(\mathbf{X}) = -\int p_{\mathbf{X}}(\mathbf{x}) \log\left(p_{\mathbf{X}}(\mathbf{x})\right) d\mathbf{x}$.

Given the above background, the concept of expected entropy decrease (expected information gain) is applied to parameter estimation of dynamical systems in the next section.

3. Parameter Information gain in dynamical systems through observations

Consider a dynamical system of the following form

$$\dot{\mathbf{x}}(t) = \boldsymbol{\mathcal{F}}\left(\mathbf{x}(t), t, \boldsymbol{\vartheta}\right), \tag{15}$$

where $\mathbf{x} \in \mathbb{R}^{N_x}$ is the state of the system, $\boldsymbol{\mathcal{F}} \in \mathbb{R}^{N_x}$ is a non-linear Lipschitz vector function, and $\boldsymbol{\vartheta} \in \mathbb{R}^{N_\vartheta}$ are the parameters. Equation (15) can be considered, for instance, as a space discretization of a non-linear PDE. An observable for the system is a vector $\mathbf{z} \in \mathbb{R}^{N_z}$ such that:

$$\mathbf{z}(t) = \mathcal{G}(\mathbf{x}(t)) + \mathbf{n}(t), \tag{16}$$

where $\mathcal{G}$ is a non-linear function, often known as the observation operator, and $\mathbf{n}(t) \in \mathbb{R}^{N_z}$ is the noise component. The goal of an inverse problem for this system is to determine the parameter vector $\boldsymbol{\vartheta}$ by measurements of $\mathbf{z}(t)$.

In equation (16) the noise is considered to be additive and i.i.d at all times. This, however, as will be presented later, is not a requirement of the presented method. The noise can also present itself in the forward model of equation (15). Furthermore, the presented method is not restricted to Gaussian nature of noise; any statistical process of noise, as long as one can draw independent samples from it, can be used.

The random variable for the parameter vector is denoted by $\boldsymbol{\Theta}$ and its (prior) probability distribution by $p_{\boldsymbol{\Theta}}(\boldsymbol{\vartheta})$. By virtue of the parameters and initial conditions for the forward model, equation (15), being random variables the state vector and the observation vectors too are random variables. The random variables for the state vector and the observation vector at time t are denoted by $\mathbf{X}(t)$ and $\mathbf{Z}(t)$, respectively. Consequently, in equations (15) and (16), the variables $\mathbf{x}(t)$, $\boldsymbol{\vartheta}$, $\mathbf{z}(t)$, and $\mathbf{n}(t)$ are realisations of the corresponding random variables $\mathbf{X}(t)$, $\boldsymbol{\Theta}$, $\mathbf{Z}(t)$, and $\mathbf{N}(t)$ (the noise process), respectively. The question of interest in this article is to estimate how much information about the parameters, $\boldsymbol{\Theta}$, is contained in a series of observations of $\mathbf{Z}(t)$.

Let $\mathbf{Z}_j$ denote the random observation vector at time t_j, and $\mathbf{z}_j \in \mathbb{R}^{N_z}$ denote the measurement of this random vector (one realisation). Similarly, let $\mathbf{Z}_{1:n}$ denote the set of n observation vectors at discrete time instants labelled t_1 to t_n, and $\mathbf{z}_{1:n} \in \mathbb{R}^{N_z \times n}$ denote the corresponding realisations. Starting from a prior probability distribution of the parameters $p_{\boldsymbol{\Theta}}(\boldsymbol{\vartheta})$, the posterior probability distribution of the parameters evolves as follows by successive observations $\mathbf{Z}_j = z_j$

$$p_{\boldsymbol{\Theta}}(\boldsymbol{\vartheta}) \to p_{\boldsymbol{\Theta}|\mathbf{Z}_1}(\boldsymbol{\vartheta}|\mathbf{z}_1) \to p_{\boldsymbol{\Theta}|\mathbf{Z}_{1:2}}(\boldsymbol{\vartheta}|\mathbf{z}_{1:2}) \to \ldots \to p_{\boldsymbol{\Theta}|\mathbf{Z}_{1:j}}(\boldsymbol{\vartheta}|\mathbf{z}_{1:j}) \to \ldots \to p_{\boldsymbol{\Theta}|\mathbf{Z}_{1:n}}(\boldsymbol{\vartheta}|\mathbf{z}_{1:n})$$

For the above probability distributions, the differential entropy evolves as

$$H(\boldsymbol{\Theta}) \to H(\boldsymbol{\Theta}|\mathbf{Z}_1) \to H(\boldsymbol{\Theta}|\mathbf{Z}_{1:2}) \to \ldots \to H(\boldsymbol{\Theta}|\mathbf{Z}_{1:j}) \to \ldots \to H(\boldsymbol{\Theta}|\mathbf{Z}_{1:n})$$

where, as before for an *a priori* analysis, the conditional differential entropies $H(\boldsymbol{\Theta}|\mathbf{Z}_{1:j})$ are averaged quantities, see equation (11), over all realisations of $\mathbf{Z}_{1:j}$

$$H(\boldsymbol{\Theta}|\mathbf{Z}_{1:j}) = \int_{\boldsymbol{\mathcal{Z}}_{1:j}} \int_{\boldsymbol{\mathcal{T}}} p_{\boldsymbol{\Theta}|\mathbf{Z}_{1:j}}(\boldsymbol{\vartheta}|\mathbf{z}_{1:j}) \log\left(\frac{1}{p_{\boldsymbol{\Theta}|\mathbf{Z}_{1:j}}(\boldsymbol{\vartheta}|\mathbf{z}_{1:j})}\right) p_{\mathbf{Z}_{1:j}}(\mathbf{z}_{1:j})\, d\boldsymbol{\vartheta}\, d\mathbf{z}_{1:j}, \tag{17}$$

where $\boldsymbol{\mathcal{Z}}_{1:j}$ and $\boldsymbol{\mathcal{T}}$ denote the support of $p_{\mathbf{Z}_{1:j}}(\mathbf{z}_{1:j})$ and $p_{\boldsymbol{\Theta}}(\boldsymbol{\vartheta})$, respectively. Following this, the quantity of interest – the expected gain in information– evolves as

$$0 \to \Delta H(\boldsymbol{\Theta}|\mathbf{Z}_1) \to \Delta H(\boldsymbol{\Theta}|\mathbf{Z}_{1:2}) \to \ldots \to \Delta H(\boldsymbol{\Theta}|\mathbf{Z}_{1:j}) \to \ldots \to \Delta H(\boldsymbol{\Theta}|\mathbf{Z}_{1:n})$$

where each $\Delta H(\boldsymbol{\Theta}|\mathbf{Z}_{1:j}) = I_{\boldsymbol{\Theta}|\mathbf{Z}_{1:j}} = M(\boldsymbol{\Theta}; \mathbf{Z}_{1:j})$ is the difference

$$\Delta H(\boldsymbol{\Theta}|\mathbf{Z}_{1:j}) = I_{\boldsymbol{\Theta}|\mathbf{Z}_{1:j}} = M(\boldsymbol{\Theta}; \mathbf{Z}_{1:j}) = H(\boldsymbol{\Theta}) - H(\boldsymbol{\Theta}|\mathbf{Z}_{1:j}). \tag{18}$$

While each of the above expected gains in information is non-negative (see remark 2.3), the evolution of expected gains in information as successive observations are added is also non-negative.

Proposition 3.1. *$H(\boldsymbol{\Theta}|\mathbf{Z}_{1:j})$ is non-increasing for increasing values of j.*

Proof. Consider the difference between two successive conditional entropies $H(\boldsymbol{\Theta}|\mathbf{Z}_{1:j+1})$ and $H(\boldsymbol{\Theta}|\mathbf{Z}_{1:j})$

$$\mathcal{D} = H(\boldsymbol{\Theta}|\mathbf{Z}_{1:j+1}) - H(\boldsymbol{\Theta}|\mathbf{Z}_{1:j}) \tag{19}$$

From equation (17), $\mathcal{D}$ can be written as

$$\mathcal{D} = \int_{\boldsymbol{\mathcal{Z}}_{1:j+1}} \int_{\boldsymbol{\mathcal{T}}} p_{\boldsymbol{\Theta},\mathbf{Z}_{1:j+1}}(\boldsymbol{\vartheta}, \mathbf{z}_{1:j+1}) \log \left(\frac{p_{\boldsymbol{\Theta}|\mathbf{Z}_{1:j}}(\boldsymbol{\vartheta}|\mathbf{z}_{1:j})}{p_{\boldsymbol{\Theta}|\mathbf{Z}_{1:j+1}}(\boldsymbol{\vartheta}|\mathbf{z}_{1:j+1})} \right) \, d\boldsymbol{\vartheta} \, d\mathbf{z}_{1:j+1} \tag{20}$$

$$\mathcal{D} = \int_{\boldsymbol{\mathcal{Z}}_{1:j+1}} \int_{\boldsymbol{\mathcal{T}}} p_{\boldsymbol{\Theta},\mathbf{Z}_{1:j+1}}(\boldsymbol{\vartheta}, \mathbf{z}_{1:j+1}) \log \left(\frac{p_{\boldsymbol{\Theta}|\mathbf{Z}_{1:j}}(\boldsymbol{\vartheta}|\mathbf{z}_{1:j}) \; p_{\mathbf{Z}_{1:j+1}}(\mathbf{z}_{1:j+1})}{p_{\boldsymbol{\Theta},\mathbf{Z}_{1:j+1}}(\boldsymbol{\vartheta}, \mathbf{z}_{1:j+1})} \right) \, d\boldsymbol{\vartheta} \, d\mathbf{z}_{1:j+1} \tag{21}$$

Since $\log(x)$ is a concave function, the inequality $\mathbb{E}[\log(x)] \leq \log(\mathbb{E}[x])$, where $\mathbb{E}[\cdot]$ denotes expectation, holds due to Jensen [26]. Applying this inequality to the expression for $\mathcal{D}$ in the above equation yields

$$\mathcal{D} \leq \log \left(\int_{\boldsymbol{\mathcal{Z}}_{1:j+1}} \int_{\boldsymbol{\mathcal{T}}} p_{\boldsymbol{\Theta},\mathbf{Z}_{1:j+1}}(\boldsymbol{\vartheta}, \mathbf{z}_{1:j+1}) \left(\frac{p_{\boldsymbol{\Theta}|\mathbf{Z}_{1:j}}(\boldsymbol{\vartheta}|\mathbf{z}_{1:j}) \; p_{\mathbf{Z}_{1:j+1}}(\mathbf{z}_{1:j+1})}{p_{\boldsymbol{\Theta},\mathbf{Z}_{1:j+1}}(\boldsymbol{\vartheta}, \mathbf{z}_{1:j+1})} \right) d\boldsymbol{\vartheta} \, d\mathbf{z}_{1:j+1} \right) \tag{22}$$

$$\mathcal{D} \leq \log \left(\int_{\boldsymbol{\mathcal{Z}}_{1:j+1}} \int_{\boldsymbol{\mathcal{T}}} p_{\boldsymbol{\Theta}|\mathbf{Z}_{1:j}}(\boldsymbol{\vartheta}|\mathbf{z}_{1:j}) \; p_{\mathbf{Z}_{1:j+1}}(\mathbf{z}_{1:j+1}) \; d\boldsymbol{\vartheta} \, d\mathbf{z}_{1:j+1} \right)$$

$$\mathcal{D} \leq \log \left(\int_{\boldsymbol{\mathcal{Z}}_{1:j}} \int_{\boldsymbol{\mathcal{T}}} p_{\boldsymbol{\Theta},\mathbf{Z}_{1:j}}(\boldsymbol{\vartheta}, \mathbf{z}_{1:j}) \; d\boldsymbol{\vartheta} \, d\mathbf{z}_{1:j} \right)$$

$$\mathcal{D} \leq \log(1) = 0 \qquad \blacksquare$$

A direct consequence of the above proposition is that the expected gain in information represented by equation (18) is non-decreasing as more (increasing j) observations are available. Intuitively, this is reasonable as the addition of more observations should not result in a decrease in the information that has already been gained. It is also reasonable that $\mathcal{D}$ in equation (19) should be zero if and only if the observation $\mathbf{Z}_{j+1}$ is totally uncorrelated to the parameter $\boldsymbol{\Theta}$ given the observations $\mathbf{Z}_{1:j}$. Alternatively, once $\mathbf{Z}_{1:j}$ are observed, if there remains no dependence between the random variables $\mathbf{Z}_{j+1}$ and $\boldsymbol{\Theta}$, then no information can be gained about the parameter by additionally observing $\mathbf{Z}_{j+1}$.

Remark 3.1. The equality in equation (22) occurs when the numerator and denominator in the log term of equation (21) are identical, i.e. when

$$p_{\boldsymbol{\Theta},\mathbf{Z}_{1:j+1}}(\boldsymbol{\vartheta}, \mathbf{z}_{1:j+1}) = p_{\boldsymbol{\Theta}|\mathbf{Z}_{1:j}}(\boldsymbol{\vartheta}|\mathbf{z}_{1:j}) \; p_{\mathbf{Z}_{1:j+1}}(\mathbf{z}_{1:j+1}) \tag{23}$$

$$= p_{\boldsymbol{\Theta}|\mathbf{Z}_{1:j}}(\boldsymbol{\vartheta}|\mathbf{z}_{1:j}) \; p_{\mathbf{Z}_{1:j}|\mathbf{Z}_{j+1}}(\mathbf{z}_{1:j}|\mathbf{z}_{j+1}) \; p_{\mathbf{Z}_{j+1}}(\mathbf{z}_{j+1}) \tag{24}$$

$$= p_{\boldsymbol{\Theta}}(\boldsymbol{\vartheta}) \; p_{\mathbf{Z}_{1:j}|\boldsymbol{\Theta}}(\mathbf{z}_{1:j}|\boldsymbol{\vartheta}) \; p_{\mathbf{Z}_{j+1}|\mathbf{Z}_{1:j}}(\mathbf{z}_{j+1}|\mathbf{z}_{1:j}). \tag{25}$$

Equations (24) and (25) show the familiar factorisation of the joint probability distribution of a Markov chain. Hence, if a Markov chain of the form $\boldsymbol{\Theta} \to \mathbf{Z}_{1:j} \to \mathbf{Z}_{j+1}$ is formed between the random variables $\boldsymbol{\Theta}$, $\mathbf{Z}_{1:j}$, and $\mathbf{Z}_{j+1}$, then there is no information gain about $\boldsymbol{\Theta}$ by observing $\mathbf{Z}_{j+1}$ when $\mathbf{Z}_{1:j}$ have already been observed. One can also view this result in the form of the Markov property that conditioned on $\mathbf{Z}_{1:j}$ the variables $\boldsymbol{\Theta}$ and $\mathbf{Z}_{j+1}$ are independent and hence carry no information about each other.

Remark 3.2. The quantity of interest presented in equation (18) can also be written as

$$\Delta H(\boldsymbol{\Theta}|\mathbf{Z}_{1:j}) = M(\boldsymbol{\Theta}; \mathbf{Z}_{1:j}) = H(\boldsymbol{\Theta}) + H(\mathbf{Z}_{1:j}) - H(\boldsymbol{\Theta}, \mathbf{Z}_{1:j}). \tag{26}$$

where

$$H(\boldsymbol{\Theta}, \mathbf{Z}_{1:j}) = \int_{\boldsymbol{\mathcal{Z}}_{1:j}} \int_{\boldsymbol{\mathcal{T}}} p_{\boldsymbol{\Theta},\mathbf{Z}_{1:j}}(\boldsymbol{\vartheta}, \mathbf{z}_{1:j}) \log \left(\frac{1}{p_{\boldsymbol{\Theta},\mathbf{Z}_{1:j}}(\boldsymbol{\vartheta}, \mathbf{z}_{1:j})} \right) \, d\boldsymbol{\vartheta} \, d\mathbf{z}_{1:j} \tag{27}$$

is the joint entropy of the parameters and the observables, and

$$H(\mathbf{Z}_{1:j}) = \int_{\boldsymbol{\mathcal{Z}}_{1:j}} p_{\mathbf{Z}_{1:j}}(\mathbf{z}_{1:j}) \log \left(\frac{1}{p_{\mathbf{Z}_{1:j}}(\mathbf{z}_{1:j})} \right) p_{\mathbf{Z}_{1:j}}(\mathbf{z}_{1:j}) \;\; d\mathbf{z}_{1:j} \tag{28}$$

and

$$H(\boldsymbol{\Theta}) = \int_{\boldsymbol{\mathcal{T}}} p_{\boldsymbol{\Theta}}(\boldsymbol{\vartheta}) \log \left(\frac{1}{p_{\boldsymbol{\Theta}}(\boldsymbol{\vartheta}} \right) p_{\boldsymbol{\Theta}}(\boldsymbol{\vartheta}) \; d\boldsymbol{\vartheta} \tag{29}$$

are the marginal entropies of the observations and the parameters, respectively. The structure of equation (26) is useful as it does not contain any conditional entropy terms and all the quantities in the RHS have the same functional form, namely that they are pure continuous Shannon entropies of the joint random variable $[\boldsymbol{\Theta}, \mathbf{Z}_{1:j}]$ and the marginal random variables $\boldsymbol{\Theta}$ and $\mathbf{Z}_{1:j}$. Consequently, if one has an estimator for continuous Shannon entropy of a random variable, then the expected gain in information can be evaluated using equation (26).

Remark 3.3. Although $\boldsymbol{\Theta}$ in equation (26) represents the random vector corresponding to all the parameters, $\boldsymbol{\Theta} = [\Theta^{(1)}, \Theta^{(2)}, \ldots, \Theta^{(N_\vartheta)}]$, of the dynamical system represented by equation (15), it can be used to calculate the expected gains in information for any combinations of parameters. For example, expected information gains for an individual parameters $\Theta^{(l)}$ is given by equation (26) when $\boldsymbol{\Theta}$ is replaced by $\Theta^{(l)}$. Similarly, expected information gain of two parameters $\Theta^{(l)}$ and $\Theta^{(m)}$ considered together is given by $\Delta H(\Theta^{(l)}, \Theta^{(m)}|\mathbf{Z}_{1:j}) = H(\Theta^{(l)}, \Theta^{(m)}) + H(\mathbf{Z}_{1:j})) - H(\Theta^{(l)}, \Theta^{(m)}, \mathbf{Z}_{1:j})$.

4. From expected information gain to assessing identifiability

Consider a single parameter $\Theta^{(l)}$ of the dynamical system. From section 3, the expected gain in information about this parameter by observing $\mathbf{Z}_{1:j}$ is $\Delta H(\Theta^{(l)}|\mathbf{Z}_{1:j})$, where $1 \leq j \leq n$ represents the time when the observation $\mathbf{Z}_j$ is taken. This section deals with interpretation of the expected gain in information in relation to identifiability of the parameter $\Theta^{(l)}$. It is assumed that one can numerically estimate this information gain

(such methods are presented in section 6). An intuitive interpretation of the expected gain in information relies on its magnitude. One would expect high $\Delta H(\Theta^{(l)}|\mathbf{Z}_{1:j})$ to imply that a substantial amount of information about the parameter $\Theta^{(l)}$ is contained in the observations $\mathbf{Z}_{1:n}$, and hence $\Theta^{(l)}$ to be identifiable. One would also expect the unidentifiable parameters to show little to no expected gain in information. Furthermore, the time-intervals $t_{a:b}$ (corresponding to $\mathbf{Z}_{a:b}$) where $\Delta H(\Theta^{(l)}|\mathbf{Z}_{1:j})$ shows most increase should be the time-intervals where the measurements of $\mathbf{Z}_{a:b}$ are most informative about the parameter. This would hint that addition of more observations in this time-interval can lead to a better estimate of the parameter.

While the evolution of $\Delta H(\Theta^{(l)}|\mathbf{Z}_{1:j})$ certainly conveys how easy or difficult (based on its magnitude) it is to identify a particular parameter $\Theta^{(l)}$ from observations of $\mathbf{Z}$, it is relatively hard to make physical sense of a statement such as *"by observing $\mathbf{Z}_{1:n}$ at times $t_{1:n}$, the expected gain in information for a particular parameter is q nats"*. Unless q takes extremely high (or extremely low) values relative to the prior entropy $H(\Theta^{(l)})$, thereby implying certain identifiability (or unidentifiability), the magnitude of q nats in itself has little physical interpretation in terms of whether the parameter is identifiable or not. Indeed, if two parameters have similar prior entropies and resulted in q and r nats of expected information gain, the relative magnitudes of q and r can be used to compare the identifiability of one parameter against the other, i.e. to ascertain how easy/difficult it is to identify one parameter with respect to the other.

In order to assign a physical interpretation to the expected information gain in relation to identifiability, a concept based on the intuitive and physically interpretable argument of direct observation is introduced. The following hypothetical question is posed:

> *Suppose there existed a hypothetical device that could directly measure the parameter Θ, i.e. it measures the parameter corrupted by an additive Gaussian noise*
>
> $$\mathfrak{Z} = \Theta + \Lambda \qquad\qquad \Lambda \sim \mathcal{N}(0, \sigma_e^2) \tag{30}$$
>
> *where $\mathfrak{Z}$ represents the hypothetical measurement and σ_e^2 represents the variance of zero-mean noise Λ in the hypothetical device. Given the gain in information that we expect from our complex forward and observation processes, what is the σ_e^2 that would result in precisely the same expected gain in information for a single measurement through the hypothetical device? .*

The hypothetical noise variance σ_e^2 is referred as the *entropy equivalent variance of a single direct observation* (EEV) and represents the precision of the hypothetical instrument that is used to directly measure the individual parameter Θ. It converts the hard-to-interpret nats of information to the variance of a measurement device with Gaussian error statistics. Since, it converts the complex forward and observation operators – that hide the manner in which information is gained about the parameters from the observables – to a simple method of directly measuring the parameter only once, it is readily interpretable.

EEV is meaningful as it offers an alternate/equivalent measure for interpreting nats of expected information gain. For instance, if the initial (prior) parameter variance is 10^{-1} square-units and σ_e^2 for this parameter is 10^2 square-units, then the interpretation is that if one were to measure this parameter directly with a hypothetical measurement device, the error in such measurement would be ± 20.0 ($\pm$ two standard deviations) units. Comparing this with the error in prior knowledge of $\pm$ 0.6 units, one may conclude that such an instrument is clearly futile to identify the parameter. Consequently, by the equivalence of entropy between the real measurements and the hypothetical measurements, the parameter can be deemed unidentifiable. On the other hand if σ_e^2 is 10^{-3} square-units then the hypothetical instrument's measurement precision is ± 0.06 units and in this case the parameter can be deemed identifiable. In what follows the formulation of σ_e^2 is presented.

In this formulation, the entropies of the hypothetical devices are represented by $\mathcal{H}$ and the entropies of the real observation process are represented by H. Consider a parameter ϑ, the corresponding random variable Θ, and related observables $\mathbf{Z}_{1:n}$ measured at times $t_{1:n}$. Assume that the real expected gain in information, $\Delta H(\Theta|\mathbf{Z}_{1:n})$, has been estimated (method presented in section 6). Let the prior entropy Θ be $H(\Theta)$. If the prior probability distribution of this parameter were Gaussian with mean μ_0 and variance σ_0^2

$$p_\Theta(\vartheta) = \mathcal{N}(\vartheta : \mu_0,\, \sigma_0^2) = \frac{1}{\sigma_0\sqrt{2\pi}} \exp\left(-\frac{(\vartheta-\mu_0)^2}{2\sigma_0^2}\right), \tag{31}$$

then its entropy is given by $\mathcal{H}(\Theta) = \frac{1}{2}\log(2\,\pi\,e\,\sigma_0^2)$. This can be equated to $H(\Theta)$ in order to obtain an entropy equivalent σ_0^2

$$\mathcal{H}(\Theta) = \frac{1}{2}\log\left(2\,\pi\,e\,\sigma_0^2\right) = H(\Theta) \tag{32}$$

Consider the direct observation process for Θ presented in equation (30). By Bayes' rule the posterior probability distribution of Θ by observing $\mathfrak{Z} = \mathfrak{z}$ is

$$p_{\Theta|\mathfrak{Z}}(\vartheta|\mathfrak{z}) = \frac{p_{\mathfrak{Z}|\Theta}(\mathfrak{z}|\vartheta)\;\; p_\Theta(\vartheta)}{p_{\mathfrak{Z}}(\mathfrak{z})}. \tag{33}$$

From equations (30) and (31) it is easy to see that the above conditional distribution is Gaussian

$$p_{\Theta|\mathfrak{Z}}(\vartheta|\mathfrak{z}) = \mathcal{N}\left(\Theta :\; \mu_1,\; \sigma_1^2\right), \tag{34}$$

where

$$\mu_1 = \frac{\sigma_e^2}{\sigma_0^2+\sigma_e^2}\;\mu_0 + \frac{\sigma_0^2}{\sigma_0^2+\sigma_e^2}\;\mathfrak{z}, \tag{35}$$

and

$$\frac{1}{\sigma_1^2} = \frac{1}{\sigma_0^2} + \frac{1}{\sigma_e^2}. \tag{36}$$

The entropy of this posterior distribution is

$$\mathcal{H}\left(\Theta|\ \mathfrak{Z}=\mathfrak{z}\right)=\frac{1}{2}\log\left(2\,\pi\,e\,\sigma_1^2\right). \tag{37}$$

Since the above conditional entropy does not depend on the value $\mathfrak{z}$ of the observation, the average conditional entropy (expectation over all values of $\mathfrak{Z}$) is identical to the above

$$\mathcal{H}\left(\Theta|\ \mathfrak{Z}\right)=\int_{\mathcal{Z}} H\left(\Theta|\ \mathfrak{Z}=\mathfrak{z}\right)p_{\mathfrak{Z}}(\mathfrak{z})\ d\mathfrak{z}=\frac{1}{2}\log\left(2\,\pi\,e\,\sigma_1^2\right). \tag{38}$$

Hence, the decrease in entropy (increase in information) from equations (32) and (38) is

$$\Delta\mathcal{H}(\Theta|\mathfrak{Z})=\mathcal{H}(\Theta)-\mathcal{H}(\Theta|\mathfrak{Z})=\frac{1}{2}\log\left(\frac{\sigma_0^2}{\sigma_1^2}\right) \tag{39}$$

Substituting σ_1^2 from equation (36) in the above equation yields

$$\Delta\mathcal{H}(\Theta|\mathfrak{Z})=\frac{1}{2}\log\left(1+\frac{\sigma_0^2}{\sigma_e^2}\right) \tag{40}$$

In order to obtain the same entropy decrease in the hypothetical measurement and the real measurements, $\Delta\mathcal{H}(\Theta|\mathfrak{Z})$ must be equal to $\Delta H(\Theta|\mathbf{Z}_{1:n})$, which yields

$$\sigma_e^2=\frac{\sigma_0^2}{\exp\left\{2\ \Delta H(\Theta|\mathbf{Z}_{1:n})\right\}-1} \tag{41}$$

Equation (41) depends on the expected gain in information and the equivalent prior variance that depends on the prior entropy, $H(\Theta)$, of the parameter through equation (32). It is interesting to see from the above equation that the following ratio

$$\sigma_u^2=\frac{\sigma_e^2}{\sigma_0^2}=\frac{1}{\exp\left\{2\ \Delta H(\Theta|\mathbf{Z}_{1:n})\right\}-1} \tag{42}$$

depends only on the information gain by measuring $\mathbf{Z}_{1:n}$. When the prior variance σ_0^2 is unity then $\sigma_u^2=\sigma_e^2$, and is therefore termed *the entropy equivalent variance of a single direct measurement with respect to unit prior variance* (EEVU). This is a standardised/normalised measure which enables easy comparison of equivalent variances between different parameters irrespective of their prior entropy magnitudes. It is meaningful due to the property of invariance of the mutual information (also the expected gain in information) under homeomorphic transformations. Following remark 2.4, since the expected gain in information remains invariant under homeomorphic transformations, one can imagine that for each parameter Θ there exists such a transformation under which the prior distribution is transformed to a distribution with unit σ_0^2. While the existence of such a transformation is a hypothesis, the interpretation of σ_u^2 under such a transformation standardises the concept of equivalent variances. To conclude this concept, irrespective of the prior distributions chosen for the parameters – which might differ both in form and location – σ_u^2 represents the error of the hypothetical measurement device under an assumption that all the prior distributions have been transformed to distributions with unit σ_0^2.

Remark 4.1. It should be noted that σ_e^2 in the above is the variance of the Gaussian-error measurement, and not the posterior expected variance of the parameters. The latter, denoted by σ_f^2, is given by $\sigma_f^2 = \sigma_e^2\sigma_0^2/(\sigma_e^2 + \sigma_0^2)$, see equation (36).

4.1. Noise process of choice

In the above section the choice of Gaussian noise for the hypothetical measurement device is purely a result of the universal ease of interpretability of the Gaussian distribution. In cases where both the prior and posterior distributions are strongly non-Gaussian, the measurement error of the hypothetical device can be described by user's choice of noise process. Let the hypothetical observation process be

$$\mathfrak{Z} = \Theta + \Lambda(\boldsymbol{\phi}) \tag{43}$$

where $\boldsymbol{\phi}$ are the parameters of the noise process that one wishes to infer. In the above, additivity of noise process is not necessary; any transformation of Θ to $\mathfrak{Z}$ through noise $\Lambda(\boldsymbol{\phi})$ can be used. The argument is same as before: what are the parameters $\boldsymbol{\phi}$ of the observation process $\Lambda(\boldsymbol{\phi})$ that result in the same gain in information as by measurement of the real observables $\mathbf{Z}_{1:n}$. While analytically intractable in most cases, this problem can be solved in an optimisation (minimisation) framework. The idea is that the numerical method to estimate the expected gain in information (presented later in section 6) relies only on samples of the random variables. Given a guess of the noise parameters $\boldsymbol{\phi}$, one can obtain samples of the noise process $\Lambda(\boldsymbol{\phi})$ and use equation (43) to obtain samples of $\mathfrak{Z}$ from the samples of Θ. From the samples of Θ and $\mathfrak{Z}$ the expected information gain $\Delta\mathcal{H}(\Theta|\mathfrak{Z})$ can be obtained by the same method (section 6) that is used to estimate the expected information gain $\Delta H(\Theta|\mathbf{Z}_{1:n})$ of the real observation process.

Our goal is to find $\boldsymbol{\phi}$ such that $\Delta\mathcal{H}(\Theta|\mathfrak{Z}(\boldsymbol{\phi})) = \Delta H(\Theta|\mathbf{Z}_{1:n})$. The optimal noise parameter vector can hence be found by minimising the L-2 norm of the difference between the real information gain $\Delta H(\Theta|\mathbf{Z}_{1:n})$ and the information gain obtained through the hypothetical noise process $\Delta\mathcal{H}(\Theta|\mathfrak{Z}(\boldsymbol{\phi}))$

$$\hat{\boldsymbol{\phi}} = \underset{\boldsymbol{\phi}}{\operatorname{argmin}} \,||\, \Delta\mathcal{H}(\Theta|\mathfrak{Z}(\boldsymbol{\phi})) - \Delta H(\Theta|\mathbf{Z}_{1:n}) \,||^2 \tag{44}$$

The above minimisation problem can be solved by any optimisation algorithm of user's choice, and is always in 2-dimensions for calculation of the expected information gain $\Delta\mathcal{H}(\Theta|\mathfrak{Z}(\boldsymbol{\phi}))$ as both Θ and $\Lambda(\boldsymbol{\phi})$ are scalar random variables. This makes the minimisation problem relatively less expensive when compared to estimating the real $\Delta H(\Theta|\mathbf{Z}_{1:n})$ in the first place (which is in $n+1$ dimensions), and offers the users to interpret the results consistent with their choice of priors and noise process.

5. Mutual information and correlation

While the analysis presented in sections 3 and 4 can be used to assess identifiability of individual parameters, it does not reveal if two or more parameter estimates

are correlated. High correlation among two or more parameters implies that these parameters can only be estimated together (not separately) [3]. This section deals with assessment of such situations. While only the case of correlation between two parameters is considered here, its extension to three or more groups of parameters is not difficult.

Suppose two parameters, $\Theta^{(l)}$ and $\Theta^{(m)}$ are individually unidentifiable based on the analysis presented in sections 3 and 4. A reasonable question then to ask is the following: *"After observing $\mathbf{Z}_{1:j}$, suppose one had an estimate – perhaps the true value or a reasonable estimate obtained through other means – of one of the parameters, say $\Theta^{(l)}$. Given that $\Theta^{(l)}$ is now known how much more information does this knowledge carry about the parameter $\Theta^{(m)}$, and how significant is this information regarding the identifiability of parameter $\Theta^{(m)}$"*. This question is answered by means of correlation between $\Theta^{(l)}$ and $\Theta^{(m)}$ and the quantity *conditional mutual information* answers precisely this question. In section 3, for individual identifiability analysis, the mutual information between each parameter and the observations is considered , see equation (18). On the contrary, in reference to the aforementioned question, the quantity of interest in this section is the conditional mutual information, i.e. mutual information between $\Theta^{(l)}$ and $\Theta^{(m)}$ given $\mathbf{Z}_{1:j}$. This quantity, denoted as $M(\Theta^{(l)};\Theta^{(m)})|\mathbf{Z}_{1:j})$, reflects the sum of two correlations: the prior correlation between $\Theta^{(l)}$ and $\Theta^{(m)}$, and the expected correlation between $\Theta^{(l)}$ and $\Theta^{(m)}$ that is developed due to the observations of $\mathbf{Z}_{1:j}$. Alternatively, having observed $\mathbf{Z}_{1:j}$, it quantifies the average amount of information that would be gained about one of the parameters by observing the other.

From equation (13), for the two random variables $\Theta^{(l)}$ and $\Theta^{(m)}$ mutual information, $M(\Theta^{(l)};\Theta^{(m)})$, is defined as follows

$$M(\Theta^{(l)};\Theta^{(m)}) = \int_{\Theta^{(m)}}\int_{\Theta^{(l)}} p_{\Theta^{(l)},\Theta^{(m)}}(\vartheta^{(l)},\vartheta^{(m)}) \log\left(\frac{p_{\Theta^{(l)},\Theta^{(m)}}(\vartheta^{(l)},\vartheta^{(m)})}{p_{\Theta^{(l)}}(\vartheta^{(l)})\, p_{\Theta^{(m)}}(\vartheta^{(m)})}\right) d\vartheta^{(l)}\, d\vartheta^{(m)} \tag{45}$$

From the above, the conditional mutual information for the observation $\mathbf{Z}_{1:j} = \mathbf{z}_{1:j}$ is

$$M(\Theta^{(l)};\Theta^{(m)}|\mathbf{Z}_{1:j} = \mathbf{z}_{1:j}) =$$
$$\int_{\Theta^{(m)}}\int_{\Theta^{(l)}} p_{\Theta^{(l)},\Theta^{(m)}|\mathbf{Z}_{1:j}}(\vartheta^{(l)},\vartheta^{(m)}|\mathbf{z}_{1:j}) \log\left(\frac{p_{\Theta^{(l)},\Theta^{(m)}|\mathbf{Z}_{1:j}}(\vartheta^{(l)},\vartheta^{(m)})|\mathbf{z}_{1:j})}{p_{\Theta^{(l)}|\mathbf{Z}_{1:j}}(\vartheta^{(l)}|\mathbf{z}_{1:j})\; p_{\Theta^{(m)}|\mathbf{Z}_{1:j}}(\vartheta^{(m)}|\mathbf{z}_{1:j})}\right) d\vartheta^{(l)}\, d\vartheta^{(m)}. \tag{46}$$

and its expected value over all all possible observations, the conditional mutual information, $M(\Theta^{(l)};\Theta^{(m)})|\mathbf{Z}_{1:j})$ is

$$M(\Theta^{(l)};\Theta^{(m)}|\mathbf{Z}_{1:j}) = \int_{\mathcal{Z}_{1:j}} M(\Theta^{(l)};\Theta^{(m)}|\mathbf{Z}_{1:j} = \mathbf{z}_{1:j})\; p_{\mathbf{Z}_{1:j}}(\mathbf{z}_{1:j})\; d\mathbf{z}_{1:j} \tag{47}$$

By methods similar to those presented in section 3, it is easy to prove that $M(\Theta^{(l)};\Theta^{(m)}|\mathbf{Z}_{1:j}) \geq 0$. In the discussion that follows, it is assumed that the prior correlation between the parameters is zero and hence $M(\Theta^{(l)};\Theta^{(m)}|\mathbf{Z}_{1:j})$ reflects purely the correlation developed between the parameters due to the observations. Consequently, if the observations of $\mathbf{Z}_{1:j}$ induce any correlation between the parameters then $M(\Theta^{(l)};\Theta^{(m)}|\mathbf{Z}_{1:j}) > 0$, while if no correlation is induced then $M(\Theta^{(l)};\Theta^{(m)}|\mathbf{Z}_{1:j}) =$

0. The former happens when the parameters $\Theta^{(l)}$ or $\Theta^{(m)}$ form a common effect for the observations $\mathbf{Z}_{1:j}$; for example, if the observable equation in terms of the parameters has a term containing a combination of the two parameters. For instance if $z(t) = g(t, \boldsymbol{\vartheta}) + f(t)/(\vartheta_l\ \vartheta_m) + \epsilon(t)$, and if the second term in the RHS is dominant then it is difficult to differentiate between θ_l and θ_m by measuring $z_{1:j}$ and consequently to identify them separately (their product, however, can be possibly identified together). Such a case is presented in section 7.1.

For increasing j, unlike the quantity $\Delta H(\Theta^{(l)}|\mathbf{Z}_{1:j}) = M(\Theta^{(l)}; \mathbf{Z}_{1:j})$ which is monotonically non-decreasing, the evolution of $M(\Theta^{(l)}; \Theta^{(m)}|\mathbf{Z}_{1:j})$, is not necessarily monotonic. The difference between two successive conditional mutual informations is given by

$$\mathcal{D}_{\mathcal{M}} = M(\Theta^{(l)}; \Theta^{(m)}|\mathbf{Z}_{1:j+1}) - M(\Theta^{(l)}; \Theta^{(m)}|\mathbf{Z}_{1:j}) \tag{48}$$

$$= M(\Theta^{(l)}; \Theta^{(m)}|\{\mathbf{Z}_{1:j}, \mathbf{Z}_{j+1}\}) - M(\Theta^{(l)}; \Theta^{(m)}|\mathbf{Z}_{1:j}), \tag{49}$$

and reflects how much more information about $\Theta^{(l)}$ can be obtained if $\mathbf{Z}_{j+1}$ was also observed after having observed both $\mathbf{Z}_{1:j}$ and $\Theta^{(m)}$. This difference can be positive, negative, or zero, and is referred as *interaction information* in literature.

The evolution of $M(\Theta^{(l)}; \Theta^{(m)}|\mathbf{Z}_{1:j})$ can reveal interesting features of the problem under consideration. The regions of time where $M(\Theta^{(l)}; \Theta^{(m)}|\mathbf{Z}_{1:j})$ is increasing reflect the time intervals where the observations of $\mathbf{Z}_j$ induce a correlation between the parameters, and the regions where $M(\Theta^{(l)}; \Theta^{(m)}|\mathbf{Z}_{1:j})$ decreases reflect the time intervals where the observations of $\mathbf{Z}_j$ destroy the hitherto built correlation between the parameters. The former happens when the observations are a *common effect* of the parameters while the latter happens when the observations are a *common cause*. When the common effect is fixed (known/observed) a correlation is induced (referred as *explaining away* in the Bayesian literature), and when the common cause if fixed, the correlation is destroyed as the observation partly explains or accounts for the hitherto built correlation. The latter can be thought of as the situation where the observation of $\mathbf{Z}_{j+1}$ makes part of the shared information between $\Theta^{(l)}$ and $\Theta^{(m)}$ (developed due to the observations of $\mathbf{Z}_{1:j}$) redundant.

Remark 5.1. In terms of the differential entropies, the conditional mutual information can be written as

$$M(\Theta^{(l)}; \Theta^{(m)}|\mathbf{Z}_{1:j}) = H(\Theta^{(l)}, \mathbf{Z}_{1:j}) + H(\Theta^{(m)}, \mathbf{Z}_{1:j}) - H(\mathbf{Z}_{1:j}) - H(\Theta^{(l)}, \Theta^{(m)}, \mathbf{Z}_{1:j}). \tag{50}$$

In light of remark 3.2, the above form is useful as it, too, contains only terms of continuous Shannon (differential) entropy. Consequently, if an estimator of differential entropy is available, then it can be used to estimate both mutual information and conditional mutual information through equations (26) and (50), respectively.

6. Numerical methods to estimate (conditional) mutual information

From sections 3 and 5, the quantities of interest to be evaluated are the mutual information, $M(\boldsymbol{\Theta}; \mathbf{Z}_{1:j}) = \Delta H(\boldsymbol{\Theta}|\mathbf{Z}_{1:j})$, and the conditional mutual information

$M(\Theta^{(l)};\Theta^{(m)}|\mathbf{Z}_{1:j})$. Furthermore, through equations (26) and (50) both these quantities can be expressed as a combination of the continuous Shannon (differential) entropies. In light of remark 5.1, in this section, numerical methods to estimate differential entropies are presented.

To reiterate, the differential entropy of a vector random variable $\mathbf{X}$ is given by

$$H(\mathbf{X}) = \int_{\boldsymbol{\mathcal{X}}} p_{\mathbf{X}}(\mathbf{x}) \ \log\left(\frac{1}{p_{\mathbf{X}}(\mathbf{x})}\right) \ d\mathbf{x}, \tag{51}$$

where $p_{\mathbf{X}}(\mathbf{x})$ is the probability density function of $\mathbf{X}$ and has a support $\boldsymbol{\mathcal{X}}$. If the analytical form of the probability density $p_{\mathbf{X}}(\mathbf{x})$ is known then one might employ numerical integration techniques to estimate the differential entropy. However, the probability distributions of interest are generally not tractable due to high non-linearity in the forward and observation operators. Furthermore, when $\mathbf{X}$ is $\mathbf{Z}_{1:j}$ then as j increases the dimension of the vector random variable increases, which makes analytical tractability impractical in most cases. Nonetheless, given the forward and observation models, *i.e.* equations (15) and (16), and prior probability distributions on the parameters and initial conditions of the forward model, one may be able to simulate a large, yet finite, number of particles (samples from the prior distribution) forward in time to obtain samples from the probability distribution of $\mathbf{Z}_{1:j}$. Hence, methods to estimate $H(\mathbf{X})$ only from samples of $\mathbf{X}$ are sought.

An overview of non-parametric estimation of entrapy estimation methods is presented in [27]. In what follows, two such concepts are briefly described. Entropy estimation methods can be broadly classified into the following categories

(i) Kernel-density estimators based: In such methods $p_{\mathbf{X}}(\mathbf{x})$ is first approximated by a kernel-density estimate (KDE). Such KDE is constructed through the samples of $\mathbf{x}$, and is used in equation (51) to evaluate the integral (either by numerical integration or Monte-Carlo estimation). See references [28, 29] for details.

(ii) Nearest neighbours based: Such estimates are based on approximating the probability density $p_{\mathbf{X}}(\mathbf{x})$ at each sample point through the distances to its nearest neighbours, and then using equation (51) to obtain a Monte-Carlo estimate of $H(\mathbf{X})$. Kozachenko and Leonenko [30] proposed such an entropy estimator estimator based on the first nearest neighbour. This estimator was extended to more robust k-nearest neighbours based estimators by Singh et. al, [31] and Kraskov et. al. [25].

In this article, the estimator proposed by Kraskov et. al. [25] after a minor modification to aid the comparison of entropies in increasing dimensions is employed. These estimators are presented next.

6.1. k-nearest neighbour based estimator by Kraskov et. al.

If one has an approximation $\hat{p}(\mathbf{x}^i)$ of the probability density $p(\mathbf{x}^i)$ then the entropy in equation (51) can be estimated using a Monte Carlo approximation as follows

$$\hat{H}(\mathbf{X}) \;=\; \frac{1}{N_{\text{ens}}}\sum_{i=1}^{N_{\text{ens}}}\log\left(\frac{1}{\hat{p}(\mathbf{x}^i)}\right) \;=\; -\frac{1}{N_{\text{ens}}}\sum_{i=1}^{N_{\text{ens}}}\log\left(\hat{p}(\mathbf{x}^i)\right), \tag{52}$$

where $\mathbf{x}^i$ represent the N_{ens} number of samples from the probability density $p(\mathbf{x})$. In k-nearest neighbour based entropy estimators the estimator $\hat{p}(\mathbf{x})$ is constructed using the k-nearest neighbours of $\mathbf{x}^i$. Kraskov et. al. [25] proposed the following estimator for $\log\left(\hat{p}(\mathbf{x}^i)\right)$

$$\log\left(\hat{p}(\mathbf{x}^i)\right) = \psi(k) - \psi(N_{\text{ens}}) - N_x\,\mathbb{E}\left[\log(\epsilon)\right] - \log(c_d), \tag{53}$$

where $\psi(\cdot)$ is the digamma function, $\mathbb{E}[\cdot]$ is the expectation operator, N_x is the dimension of $\mathbf{x}$ ($\mathbf{x}\in\mathbb{R}^{N_x}$), ϵ is twice the distance between $\mathbf{x}^i$ and its k^{th} nearest neighbour, and c_d is the volume of the N_x-dimensional unit ball. The value of c_d depends on the kind of norm used for calculating the distances between neighbours: for max norm $c_d = 1$ and for Euclidean norm $c_d = \frac{\pi^{N_x/2}}{2^{N_x}\Gamma(1+N_x/2)}$, where Γ represents the gamma function . Based on this, the estimator for $H(\mathbf{X})$ can be written from equation (52) as

$$\hat{H}(\mathbf{X}) = \psi(N_{\text{ens}}) - \psi(k) + \log(c_d) + \frac{N_x}{N_{\text{ens}}}\sum_{i=1}^{N_{\text{ens}}}\log\left(\epsilon(i)\right) \tag{54}$$

where $\epsilon(i)$ represents twice the distance from the i^{th} sample to its k^{th} neighbour.

Given the above differential entropy estimator, the mutual information between two random variables $\mathbf{X}\in\mathbb{R}^{N_x}$ and $\mathbf{Y}\in\mathbb{R}^{N_y}$ can be calculated from the samples $(\mathbf{x}^i,\mathbf{y}^i)$ is $M(\mathbf{X};\mathbf{Y}) = H(\mathbf{X}) + H(\mathbf{Y}) - H(\mathbf{X},\mathbf{Y})$. However, Kraskov et. al. pointed out that such an approach leads to biased results as the k-nearest neighbour distances are calculated in different spaces – the joint and the marginal – which use different distance scales. To remedy this, they suggested that while equation (54) can be used to estimate the entropy in the joint space, the distance scale in the joint space can be used in the marginal spaces too in order to cancel the biases. This can be achieved by calculating the k^{th}-neighbour distances $\epsilon(i)$ in the joint space and then finding the number of nearest neighbours within this distance in the marginal spaces for marginal entropy estimation. Thus, one uses a variable k in the marginal spaces based on the distance calculated in the joint space. The entropy estimators for the joint and the marginal spaces can then be written as

$$\hat{H}(\mathbf{X},\mathbf{Y}) = \psi(N_{\text{ens}}) - \psi(k) + \log(c_{d_{\mathbf{x}}}c_{d_{\mathbf{y}}}) + \frac{N_x+N_y}{N_{\text{ens}}}\sum_{i=1}^{N_{\text{ens}}}\log\left(\epsilon^J(i)\right) \tag{55}$$

and

$$\hat{H}(\mathbf{X}) = \psi(N_{\text{ens}}) - \frac{1}{N_{\text{ens}}}\sum_{i=1}^{N_{\text{ens}}}\psi\left(k_{\mathbf{x}}(i)+1\right) + \log(c_{d_{\mathbf{x}}}) + \frac{N_x}{N_{\text{ens}}}\sum_{i=1}^{N_{\text{ens}}}\log\left(\epsilon^J(i)\right) \tag{56}$$

$$\hat{H}(\mathbf{Y}) = \psi(N_{\text{ens}}) - \frac{1}{N_{\text{ens}}}\sum_{i=1}^{N_{\text{ens}}} \psi\left(k_{\mathbf{y}}(i)+1\right) + \log(c_{d_{\mathbf{y}}}) + \frac{N_y}{N_{\text{ens}}}\sum_{i=1}^{N_{\text{ens}}} \log\left(\epsilon^J(i)\right) \tag{57}$$

where for each sample point $(\mathbf{x}^i, \mathbf{y}^i)$ the distance $\epsilon^J(i)$ is twice the distance to the k-th nearest neighbour in the joint space, $k_{\mathbf{x}}(i)$ and $k_{\mathbf{y}}(i)$ represent the number of points in the marginal spaces of $\mathbf{X}$ and $\mathbf{Y}$ that are at distances less than $\epsilon^J(i)/2$ from the points $\mathbf{x}^i$ and $\mathbf{y}^i$, respectively, and $c_{d_{(\cdot)}}$ represents the volume of a unit ball in the dimension of $(\cdot)$. The mutual information estimator then simplifies to

$$\hat{M}(\mathbf{X},\mathbf{Y}) = \psi(N_{\text{ens}}) + \psi(k) - \frac{1}{N_{\text{ens}}}\sum_{i=1}^{N_{\text{ens}}} \psi\left(k_{\mathbf{x}}(i)+1\right) - \frac{1}{N_{\text{ens}}}\sum_{i=1}^{N_{\text{ens}}} \psi\left(k_{\mathbf{y}}(i)+1\right) \tag{58}$$

6.2. Modification of the estimator

As mentioned in section 3 the quantity of interest in order to quantify the information gain about the parameters is the evolution of $M(\mathbf{\Theta}; \mathbf{Z}_{1:j})$, see equation (26) for increasing j. Equations (55), (56) and (57) can be used to estimate this mutual information, equation (58), without incorporating additional biases due to different distance scales in the joint and marginal spaces. However, a comparison of $\hat{M}(\mathbf{\Theta}; \mathbf{Z}_{1:j})$ with $\hat{M}(\mathbf{\Theta}; \mathbf{Z}_{1:j+1})$ is unfair as the space $(\mathbf{\Theta}, \mathbf{Z}_{1:j})$ is of lower dimension than that of $(\mathbf{\Theta}; \mathbf{Z}_{1:j+1})$. This is particularly noticeable (see Figure 4 in section 7.1) in regions of time where no information is available about the parameters, *i.e.* regions where $M(\mathbf{\Theta}; \mathbf{Z}_{1:j})$ remains constant for changes in j. In such regions the the evolution of $\hat{M}(\mathbf{\Theta}; \mathbf{Z}_{1:j})$ can decrease for increasing j, which is contrary to the non-decreasing nature of $M(\mathbf{\Theta}; \mathbf{Z}_{1:j})$, see proposition 3.1.

To avoid the above problem a minor modification in the entropy estimator is proposed. Once the sampling plan is defined, *i.e.* once the times at which observations are available is fixed as $\{t_j\}$, $1 \leq j \leq n$, the largest joint space is of $(\mathbf{\Theta}, \mathbf{Z}_{1:n})$. In this largest space, for each sample point $(\boldsymbol{\vartheta}^i, \mathbf{z}^i_{1:n})$, let $\epsilon^M(i)$ represent twice the distance to its k-th nearest neighbour. It is proposed that to calculate the mutual information estimates of $\hat{M}(\mathbf{\Theta}; \mathbf{Z}_{1:j})$ for all $1 \leq j \leq n$, the distances used be $\epsilon^M(i)$. Following equation (58), these estimates can be written as

$$\hat{M}(\mathbf{\Theta}; \mathbf{Z}_{1:j}) = +\frac{1}{N_{\text{ens}}}\sum_{i=1}^{N_{\text{ens}}} \psi\left(k_{\boldsymbol{\vartheta},\mathbf{z}_{1:j}}(i)+1\right) - \frac{1}{N_{\text{ens}}}\sum_{i=1}^{N_{\text{ens}}} \psi\left(k_{\mathbf{z}_{1:j}}(i)+1\right) - \frac{1}{N_{\text{ens}}}\sum_{i=1}^{N_{\text{ens}}} \psi\left(k_{\boldsymbol{\vartheta}}(i)+1\right) \tag{59}$$

where for each sample point $(\boldsymbol{\vartheta}^i, \mathbf{z}^i_{1:j})$, $k_{\boldsymbol{\vartheta},\mathbf{z}_{1:j}}(i)$ represents the number of points at distances less than $\epsilon^M(i)/2$ in the joint space of $(\mathbf{\Theta}, \mathbf{Z}_{1:j})$. Similarly, in the marginal spaces of $\mathbf{\Theta}$ and $\mathbf{Z}_{1:j}$, $k_{\boldsymbol{\vartheta}}(i)$ and $k_{\mathbf{z}_{1:j}}(i)$ represent the number of points at distances less than $\epsilon^M(i)/2$ from the sample points $\boldsymbol{\vartheta}^i$ and $\mathbf{z}^i_{1:j}$, respectively. Since all the entropy estimators use the same distance scale ϵ^M, the successive mutual information estimates, $\hat{M}(\mathbf{\Theta}; \mathbf{Z}_{1:j})$, can be compared to each other. In what follows, this modified mutual information estimator is used to study the evolution of gain in parameter information as more observations are available.

Lastly, the conditional mutual information $M(\Theta^{(l)};\Theta^{(m)}|\mathbf{Z}_{1:j})$ can be estimated through equation (50) by estimating all the differential entropies on the RHS of equation (50) through the estimator of Kraskov et. al. Here too, for comparison of $M(\Theta^{(l)};\Theta^{(m)}|\mathbf{Z}_{1:j})$ for different j, the k^{th}-th neighbour distances can be calculated in the largest space of $(\Theta^{(l)},\Theta^{(m)},\mathbf{Z}_{1:n})$ and all terms on the RHS of equation (50) can be estimated by finding the number of neighbours within this distance in the marginal spaces.

6.3. The numerical method

Given a dynamical system and its numerical discretisation the proposed approach consists of the following steps:

(i) generate N_{ens} samples from the prior distributions of the parameters and initial conditions.

(ii) for each sample compute the numerical solutions.

(iii) compute the corresponding observations by generating samples from the observation noise process.

(iv) compute the (conditional) mutual informations using the modified estimator presented above.

(v) compute corresponding equivalent variances.

The pseudo-code for such computation is shown in algorithm 1. One difficulty associated with the above presented estimators is that if one of the parameter magnitudes (or one of the observables magnitude) is significantly larger than others then this parameter dominates ϵ^M. Consequently, in the marginal spaces, $k_{(\cdot)}$ can encompass a large number of sample points – sometimes even the entire ensemble – which leads to erroneous results. In order to alleviate this difficulty the following procedure is performed: since the gain in information is invariant under transformations, see remark 2.4, , scaling and centering (to zero mean and unit variance) of all the components of the parameter and observation vectors is performed before calculating $\epsilon^M(i)$ and number of points inside the ϵ-balls.

The computational features of the estimator, the convergence with respect to the number of samples and neighbours, as well as its complexity are detailed in [25]. The advantage of using a Monte Carlo approach is that despite its slow convergence rate, the convergence is independent of the dimension of the stochastic space considered. This is of the utmost relevance when considering the present application, in which the dimension is growing. Moreover, it is non-intrusive and it does not require for the direct solvers to be re-written in order to take the stochasticity into account.

Algorithm 1: Algorithm to calculate gain in information for the parameters

Input:

- The forward model: $\dot{\mathbf{x}}(t) = \mathcal{F}(\mathbf{x}(t), t, \boldsymbol{\vartheta})$; $\mathbf{x} \in \mathbb{R}^{N_x}, \boldsymbol{\vartheta} \in \mathbb{R}^{N_\vartheta}$
- The observation model: $\mathbf{z}(t) = \mathcal{G}(\mathbf{x}(t)) + \mathbf{n}(t)$; $\mathbf{z}, \mathbf{n} \in \mathbb{R}^{N_z}$
- Prior probability distribution on the parameter vector: $p_{\boldsymbol{\Theta}}(\boldsymbol{\vartheta})$.
- Prior probability distribution on the initial state vector: $p_{\mathbf{X}}(\mathbf{x})$.
- Number of samples to be simulated: N_{ens}.
- List of measurements times: t_j, $1 \leq j \leq n$.
- Number of nearest neighbours to used for mutual information estimation: k
- The set $\mathcal{S}$ containing parameter combinations for which $\Delta\hat{H}_{0\to 1:j}$ is estimated. For example, $\mathcal{S} = \{\vartheta^{(1)}, \vartheta^{(2)}, (\vartheta^{(1)}, \vartheta^{(2)})\}$

Output: $\Delta\hat{H}_{0\to 1:j}$, $1 \leq j \leq n$, for all elements of $\mathcal{S}$.

begin
  Generate N_{ens} independent samples from $p_{\mathbf{X}}(\mathbf{x})$ and $p_{\boldsymbol{\Theta}}(\boldsymbol{\vartheta})$ to create $(\mathbf{x}_0^i, \boldsymbol{\vartheta}^i); 1 \leq i \leq N_{\text{ens}}$
  for $i \leftarrow 1$ **to** N_{ens} **do** `/* easily parallelisable */`
    for $j \leftarrow 1$ **to** n **do**
      $\mathbf{x}_j^i \leftarrow$ solve $\dot{\mathbf{x}}(t) = \mathcal{F}\left(\mathbf{x}(t), t, \boldsymbol{\vartheta}^i\right)$, for $\mathbf{x}^i$ at t_j with initial value $\mathbf{x}(0) = \mathbf{x}_0^i$
    end
  end
  for $i \leftarrow 1$ **to** N_{ens} **do** `/* easily vectorised */`
    for $j \leftarrow 1$ **to** n **do**
      $\mathbf{n}_j^i \leftarrow$ Generate a sample from the noise process $\mathbf{n}(t_j)$;
      $\mathbf{z}_j^i \leftarrow \mathcal{G}(\mathbf{x}_j^i) + \mathbf{n}_j^i$
    end
  end
  Centre (subtract mean) and scale (to unit variance) the ensembles of $\boldsymbol{\theta}$ and $\mathbf{z}_j$;
  for $j \leftarrow 1$ **to** n **do** `/* easily parallelisable */`
    for $i \leftarrow 1$ **to** N_{ens} **do** $\mathbf{z}_{1:j}^i \leftarrow [\mathbf{z}_1^i, \mathbf{z}_2^i, \ldots, \mathbf{z}_j^i]$;
  end
  foreach $\boldsymbol{\theta}$ *in the set* $\mathcal{S}$ **do** `/* easily parallelisable */`
    for $i \leftarrow 1$ **to** N_{ens} **do**
      $\epsilon_{\boldsymbol{\theta}}^M(i) \leftarrow$ distance of $(\boldsymbol{\theta}^i, \mathbf{z}_{1:n}^i)$ to its k^{th} nearest neighbour
    end
  end
  for $j \leftarrow 1$ **to** n **do** `/* easily parallelisable */`
    foreach $\boldsymbol{\theta}$ *in the set* $\mathcal{S}$ **do**
      Calculate $\Delta\hat{H}_{0\to 1:j}$ using $\mathbf{z}_{1:j}^i$, $\boldsymbol{\theta}^i$, and $\epsilon_{\boldsymbol{\theta}}^M(i)$; $1 \leq i \leq N_{\text{ens}}$, and eq. (59)
    end
  end
end

7. Numerical experiments.

In this section the numerical experiments are commented, of increasing complexity, aiming at validating the proposed approach.

The first example deals with an RCR circuit, which is an example of 0-D model, often used as boundary condition for simulation of cardiovascular flows, and whose identification is a key step to the setting up of realistic simulations.

Then, an epidemiological SIR model is investigated aiming at comparing the results with those obtained in the literature by means of more complex differential algebra methods. This is still an ODE model, but it is non-linear, with a large number of parameters.

A first example with PDEs is on a potential identification problem for an hyperbolic system in $1+1$ dimensions.

In the last section, the identification of sources for a problem of advection diffusion in 2D is commented.

For all the testcases presented, the number of direct simulation computed for the MC estimation of the entropies were empirically evaluated by considering the convergence of the results. The experiments presented uses a number of particles such that the estimated quantities vary less than 0.01 when the number of particles is doubled.

The number of k-neighbours has been adjusted in an analogous way.

7.1. Three-element Windkessel model

Parameter estimation in a three-element Windkessel model [32], a commonly imposed boundary condition to study haemodynamics [33], is considered. This model, as shown in Figure 1a, consists of three elements: a proximal resistance, R_p, representing the sum of large vessels; a capacitance, C, representing the elastance of large vessels; and a distal resistance, R_d, representing the resistance of smaller vessels in the microcirculation. The behaviour of the Windkessel model of Figure 1a is governed by the following differential algebraic equations

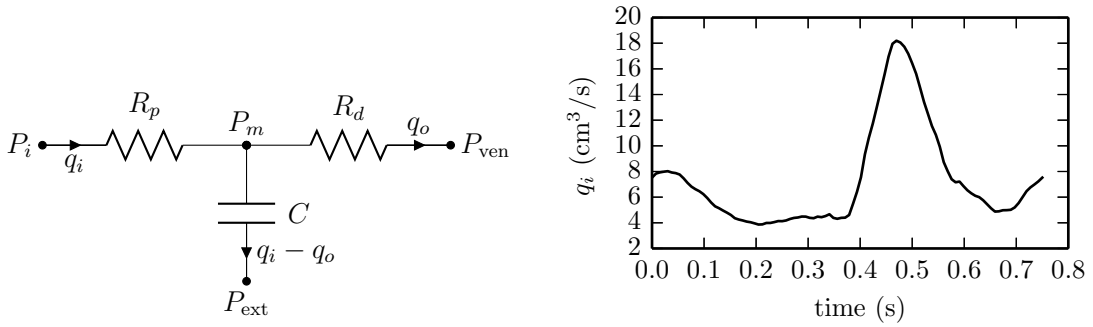
$$P_i(t) - P_m(t) = R_p\, q_i(t), \tag{60}$$

$$P_m(t) - P_{\text{ven}}(t) = R_d\, q_o(t), \tag{61}$$

$$q_i(t) - q_o(t) = C\, \frac{d}{dt}\Big(P_m - P_{\text{ext}}(t)\Big), \tag{62}$$

where q_i and q_o represent the inlet and outlet flow-rates, P_i and P_m represent the inlet and the mid-Windkessel pressure, and P_{ven} and P_{ext} represent the venous and external pressures, respectively. P_{ven} and P_{ext} are both assumed to be zero. The inlet flow-rate, the control curve, is known and is shown in Figure 1b. The goal of this analysis is to determine the identifiability of the Windkessel parameters – R_p, C, and R_d – by discrete measurements of $P_i(t)$

$$\tilde{z}(t_j) = P_i(t_j) + \tilde{\epsilon}(t_j), \tag{63}$$



(a) Schematic of a three-element Windkessel model

(b) Inlet flow-rate curve to drive the forward model

Figure 1: The three-element Windkessel model and the flow-rate curve

where $\tilde{z}(t_j)$ represents the measurement of $P_i(t_j)$ at time t_j, and $\tilde{\epsilon}(t_j)$ represents the associated measurement noise. To maintain positivity of the Windkessel parameters and pressure the following transformation is used for these variables

$$\Psi^{(\cdot)} = \Psi_0^{(\cdot)}\, 2^{\psi^{(\cdot)}} \qquad \text{where } (\cdot) \in \{P_i, R_p, C, R_d\} \tag{64}$$

where $\Psi_o^{(\cdot)}$ represent a reference value, $\Psi^{(\cdot)}$ represents the real values of P, R_p, C, R_d which can lie between 0 and ∞, and ψ represents the transformed values which can take any value between $-\infty$ and ∞. The advantage of such transformation is that the probability distributions on $\{P_i, R_p, C, R_d\}$ in the ψ-space can be assumed to have support over the entire real line while ensuring that the model is physically correct in terms of positivity of these parameters. Moreover, while the assumption of a Normal process for $\epsilon(t_i)$ in equation (63) would have been wrong by leaving a finite probability that the pressure measurement, $z(t_i)$, could be negative, the assumption of Normal noise in the measurement of the transformed variable ψ^{P_i} is not unreasonable

$$z(t_j) = \psi^{P_i(t_j)} + \epsilon(t_j), \qquad \text{where } \epsilon(t_i) = \mathcal{N}(0, \sigma_n^2). \tag{65}$$

7.1.1. Expected information gain: For the calculation of expected information gains, 150 measurements of $z(t_j)$, for t_j uniformly distributed in [0.0, 0.75] are considered. For the transformation the reference values are: $\Psi_0^{P_i} = 78.64$ mmHg; $\Psi_0^{R_p} = 0.838$ mmHg-s/cm^3; $\Psi_0^{C} = 0.0424$ cm^3/mmHg; and $\Psi_0^{R^d} = 9.109$ mmHg-s/cm^3. In the ψ-space, Normal priors of zero mean and variance 1.0 are considered for P_i, R_p, C, and R_d. The sample size is $N = 3000$ and the number of nearest neighbours is $k = 10$. Four successively increasing values of the noise variance, σ_n^2, are considered. These noise levels transform to variances of approximately 2.0, 18.0, 36.0, and 184.0 for the measurement of pressure in the real Ψ-space (a log-normal distribution when a normal distribution is considered the ψ-space).

The expected information gains for individual parameters, pairs of parameters considered together, and the triplet consisting of all the three parameters, are shown in Figure 2. The sample mean of the observations is also shown. From the relative

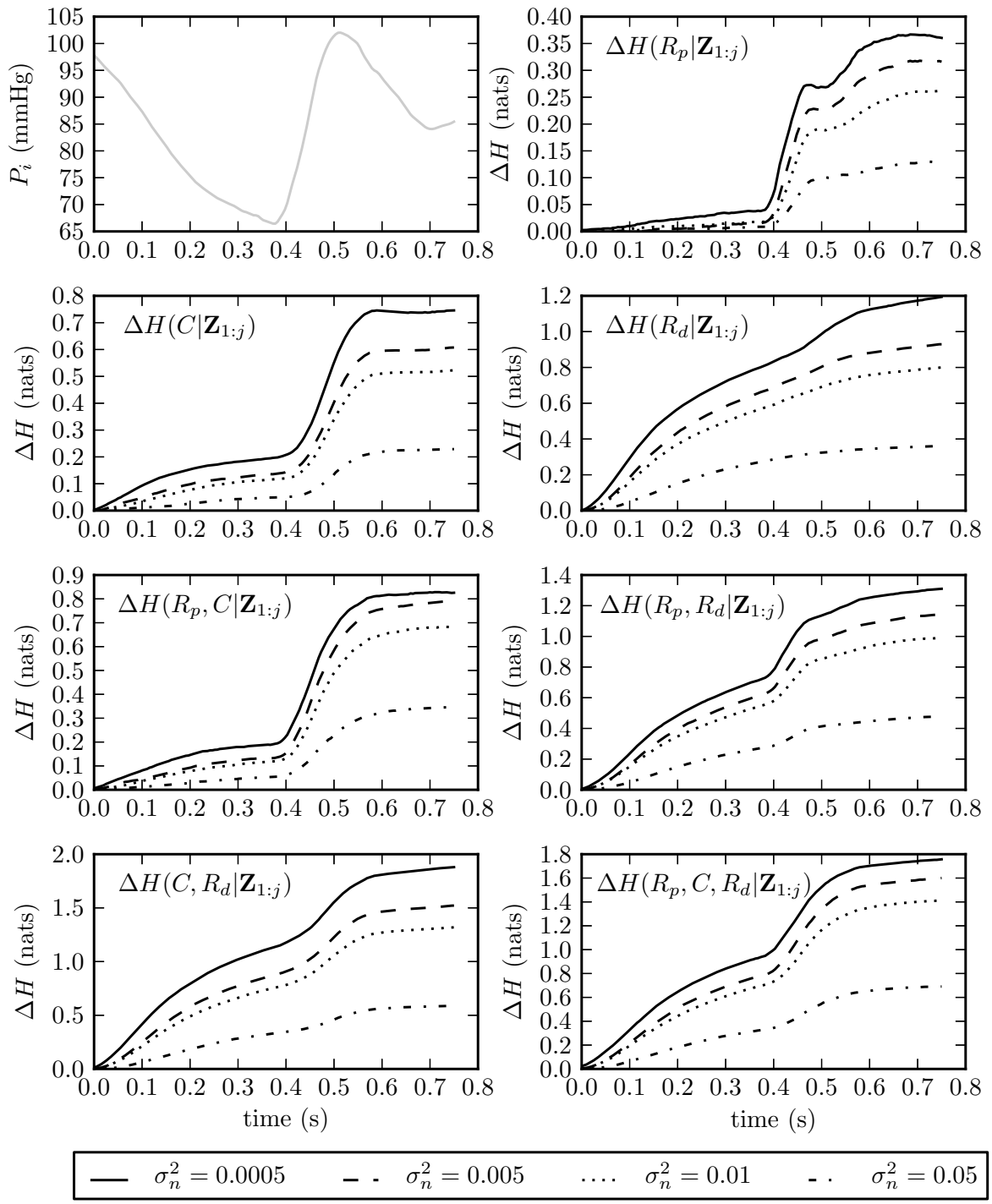


Figure 2: Gain in information for various parameter combinations in the Windkessel model. The plot for P_i represents the mean of the observations generated through the propagated samples. σ_n^2 corresponds to the transformed variable ψ^{P_i} (equation (65)). For $\sigma_n^2 = \{0.0005, 0.005, 0.01, 0.05\}$, the corresponding variances of the noise process for the real parameter P_i (untransformed, see equation (63)) are $\{1.8, 17.8, 35.7, 183.7\}$ mmHg^2.

Θ	Prior						Observation Noise		Expected Posterior	
	ψ-space		Real (Ψ-space)				ψ-space	Ψ-space		
	μ_ϑ	σ^2_ϑ	Arith. mean	Arith. var.	Geom. mean	Geom. var.	σ^2_n	σ^2_n	σ^2_e	σ^2_u
P	0.0	1.0	99.99	6167	78.64	1.62	-	-	-	-
R_p	0.0	1.0	1.065	0.7	0.838	1.62	0.0005	1.8	0.976	0.976
							0.005	17.8	1.066	1.066
							0.01	35.7	1.285	1.285
							0.05	183.7	3.426	3.426
C	0.0	1.0	0.054	0.0018	0.0424	1.62	0.0005	1.8	0.283	0.283
							0.005	17.8	0.392	0.392
							0.01	35.7	0.557	0.557
							0.05	183.7	1.747	1.747
R_d	0.0	1.0	11.58	82.7	9.109	1.62	0.0005	1.8	0.100	0.101
							0.005	17.8	0.188	0.188
							0.01	35.7	0.257	0.257
							0.05	183.7	0.912	0.912

Table 1: Real parameters are in the Ψ-space and a parameterisation of the form $\Psi = 2^\psi$ is used to ensure positivity of the parameters and pressure.

magnitudes of expected information gains for the individual parameters it can be concluded that R_d is most easily identifiable (largest gain in information) while R_p is most difficult to identify (lowest gain in information). Furthermore, it is observed that most of the information about the parameter R_p is contained in the sharp systolic phase of pressure ($t \in [0.4, 0.5]$). This is consistent with the results obtained through generalised sensitivity function (GSF) analysis presented in [34, 35]. In fact, if one normalises the expected gain in information for the individual parameters to lie between 0 and 1 then the normalised plots look similar to GSFs (see Figure 10 of [35]). While GSFs can oscillate in the presence of parameter correlations, thereby making them relatively hard to interpret, the expected information is always non-decreasing. This property makes the expected information gain easier to interpret. Furthermore, in contrast to GSFs the magnitudes of expected information gain provide a quantitative measure of not only which regions of time contain most information about a particular parameter but also how much.

The effect of increasing noise in Figure 2 is consistent with the intuition that a higher noise should result in a lesser amount of information gain. The EEV values for the parameters for the four levels of noise considered are shown in Table 1. For the lowest level of noise corresponding to noise variance of 1.8 mmHg^2 for P_i, it is observed that the σ^2_e for R_d is 0.1, implying that relative to the prior variance of 1.0 one obtains a

sharp measurement of variance 0.1 through the hypothetical measurement device. One may conclude with these statistics that the parameter R_d is easily identifiable. Similar arguments hold for the parameter C which has an $\sigma_e^2 = 0.28$, a low value compared to the prior unit variance. However, for R_p the $\sigma_e^2 = 0.97$, a value of roughly the same magnitude as the prior unit variance. This implies that even with a low level of noise, since the hypothetical measurement device has an error variance of the same amount as the prior variance, the variance of the final estimate is only halved, see equation (36). This suggests that even in the presence of low noise, the parameter R_p is hard to estimate. Furthermore, for the highest level of noise the parameter R_p has an $\sigma_e^2 = 3.4$, implying that virtually no information is contained in the measurements for this level of noise and that the parameter cannot be identified. On the other hand, the parameter R_d has an σ_e^2 of 0.91 for the highest level of noise, which is lower than the σ_e^2 of R_p for the lowest level of noise, implying that even with the highest level of noise considered, the uncertainty (judging by the variance) can be halved for R_d. Such an analysis can be used in determining the level of acceptable noise in the real measurement device when used for parameter identification. For example, consider two pressure measurement devices of variances 1.8 and 17.8 mmHg^2. If the goal of buying this measurement device was solely to estimate the Windkessel parameters, then one may conclude that the latter device, even though significantly more error-prone, will do only marginally worse than the former device when it comes to estimating R_p, C, and R_d. Consequently, in a cost-benefit analysis under the assumption that the former device is considerably expensive than the latter, the latter device should be preferred.

7.1.2. Conditional mutual information: In Figure 2 the expected gains in information when any two parameters are considered together, and when all the three parameters are considered together, are also presented. These are presented for sake of completeness; a utility of these plots lies in cases where the expected gain in information of the pair of parameters considered together is significantly larger than the sum of individual expected information gains. Such a case would imply that one can learn a lot more about the joint parameter distribution compared to the individual marginal distributions. This implies that the parameter estimates are correlated. However, since kNN estimators are used – making it unwise to quantitatively add or subtract the plots shown in Figure 2 since different length scales are used in each plot – it is proposed that that correlations be studied through the conditional mutual information estimated by equation (50) where the same distance scale is used for calculating all the RHS terms. The pair-wise conditional mutual informations are plotted in Figure 3. The first observation is that the behaviour with respect to noise is the same as for the individual information gains. This is expected. The second observation is that the conditional mutual informations for the pairs $\{R_p, C\}$ and $\{R_p, R_d\}$ are negligible, implying that these pairs of parameters are not correlated. The evolution of conditional mutual information for the pair $\{R_d, C\}$, however, shows an interesting pattern. It shows an increasing behaviour in the diastolic phase ($t \in [0.0, 0.4] \cup [0.55, 0.7]$), and a decreasing behaviour in the systolic phase

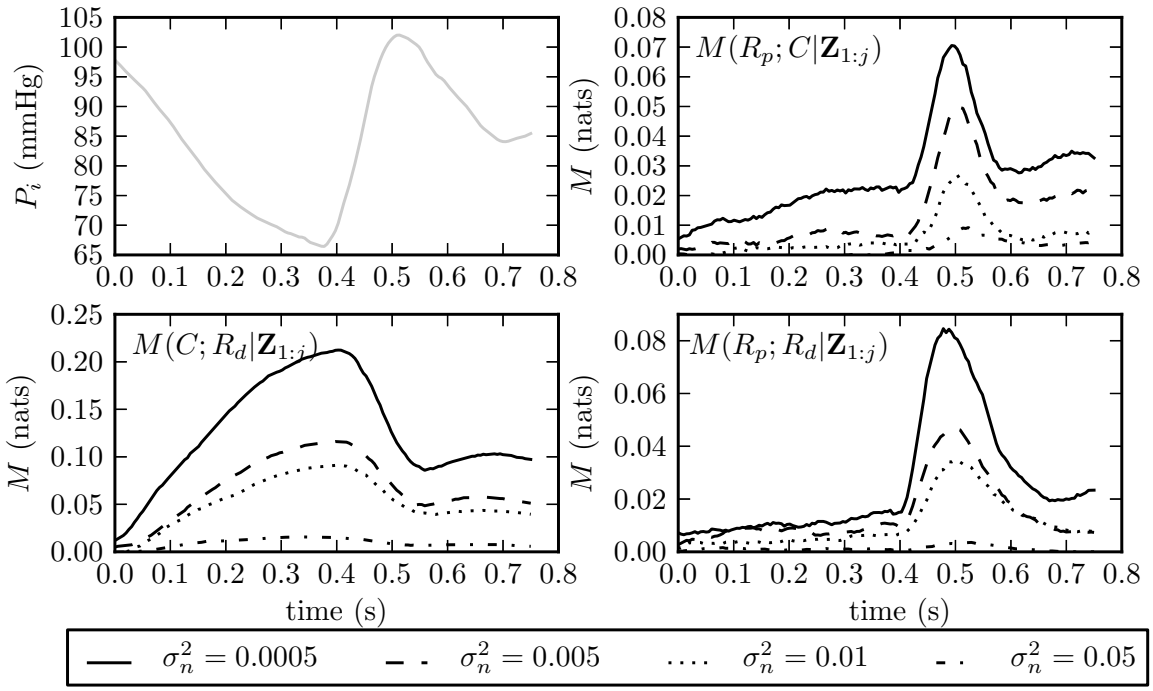


Figure 3: Mutual information between all pairs of the parameters in the Windkessel model. The plot for P_i represents the mean of the observations generated through the propagated samples. σ_n^2 corresponds to the transformed variable ψ^{P_i} (equation (65)). For $\sigma_n^2 = \{0.0005, 0.005, 0.01, 0.05\}$, the corresponding variances of the noise process for the real parameter P_i (untransformed, see equation (63)) are $\{1.8, 17.8, 35.7, 183.7\}$ mmHg^2.

($t \in [0.4, 0.55]$). From the discussion presented in section 5, regions of increasing conditional mutual information are regions where the observations build up a correlation between the parameters. This implies that the observations are largely a result of a common effect of the two parameters (hinting that a functional term containing the two parameters drives the observations) thereby building a correlation between the parameters. This makes sense if one looks at the solution to the system presented in equations (60) – (62)

$$P_i(t) = \overbrace{(P_i(0) - R_p\, q_i(0))\, e^{-t/(R_d C)}}^{\text{term-1}} + \underbrace{R_p\, q_i(t)}_{\text{term-2}} + \overbrace{e^{-t/(R_d C)}}^{\text{term-3}} \underbrace{\int_0^t \frac{e^{\tilde{t}/(R_d C)}}{C} q_i(\tilde{t}) d\tilde{t}}_{\text{term-4}} \tag{66}$$

It can be seen that solution consists of many parts where the product of the parameters R_d and C comes together as a rate of exponential decay. Their combined effect is much more dominant in the diastolic phase where term-2 and term-4 remain approximately constant (since $q_i(t)$ shows little relative variation in these regions, see Figure 1b). Consequently, in these regions, a large part of the solution, $P_i(t)$, is determined by

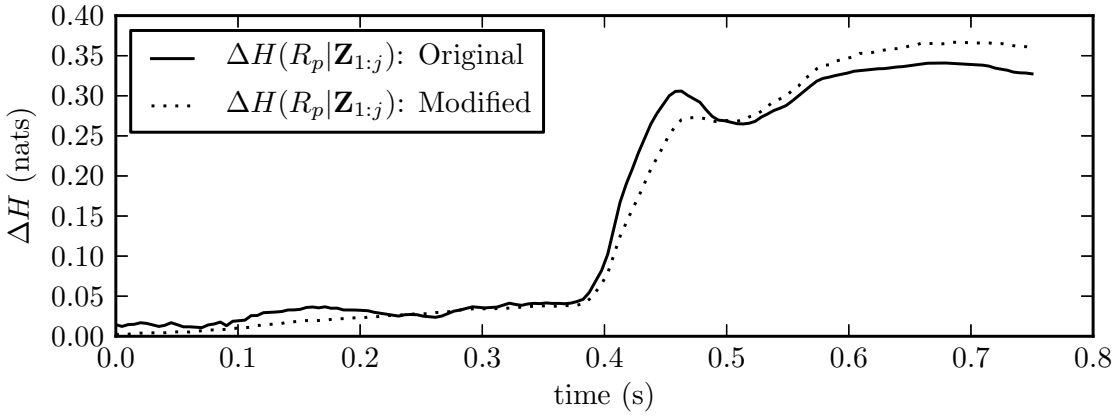


Figure 4: Comparison of the original Kraskov estimator and the modified estimator. The modified estimator handles the regions of constant entropy, $t \in [0.45, 0.5]$, better than the original estimator which shows a decrease in such regions.

the combined effect of these two parameters, which results in an induced correlation by measurements of $P_i(t)$. It is interesting to note that even if the solution in equation (66) was not known, as typically would be the case for complex dynamical systems, the conditional mutual information plots hint at which functional terms of parameter combinations appear together in the solution or influence the solution together. Furthermore, they isolate regions of time where this combined effect is most dominant. On the other hand, regions of decreasing conditional mutual information are regions where given the observations some of the information that one parameter contains about the other becomes redundant. For example in the systolic phase of pressure ($t \in [0.4, 0.55]$) the built-up correlation between the R_d and C is destroyed as observing $P_i(t)$ in this region partly accounts for the correlation. After the cycles (if any) of the increase and decrease of conditional mutual information, the value of the conditional mutual information at the end, i.e. when all measurements have been taken, is of high interest. This quantity reflects the final correlation that is induced between the parameters after all the observations have been taken into account. From Figure 3, this final value for the pair $\{R_d, C\}$ is approximately 0.1 nats for the lowest level of noise. This implies that at the end of the experiment, if one fixes/knows the value of one of the parameters, then 0.1 nats of information could be gained about the other parameter on top of the already gained information shown in Figure 2. Since this is only 12.5% more information for the parameter C and 0.8% more information for R_d, it is safe to conclude that the correlations are not significant to affect identifiability.

7.1.3. Modified Windkessel model: In order to demonstrate an example where correlations affect identifiability, a simple modification of the Windkessel model is considered where an extra distal resistance is added (see Figure 5). In this model, $R_{d_1} + R_{d_2}$ behaves as R_d of the previous model and hence individually, we expect

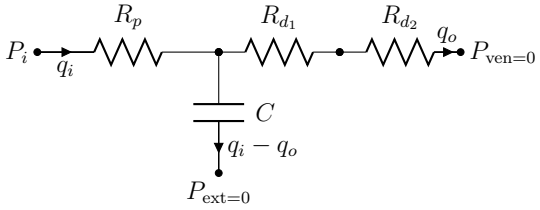


Figure 5: Schematic of a modified three-element Windkessel model

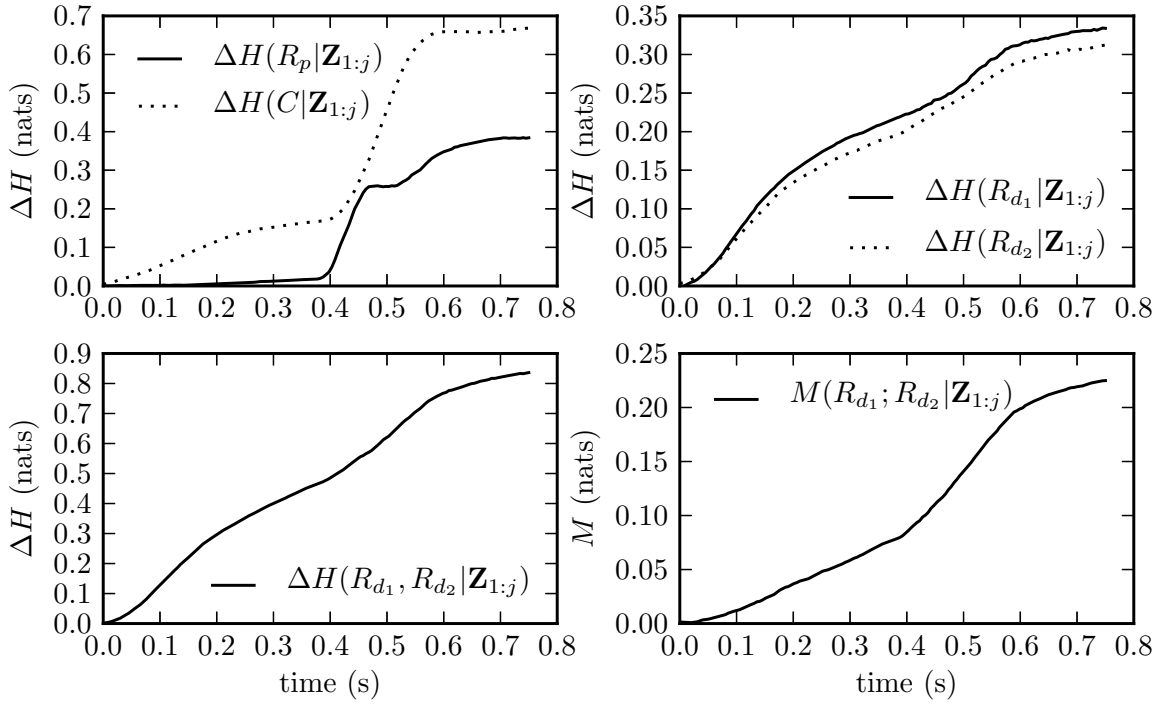


Figure 6: Mutual information between R_{d_1} and R_{d_2}

Parameter	prior variance	σ_e^2	σ_u^2	σ_u^2 with MI
R_p	1.0	0.86	0.86	
C	1.0	0.35	0.35	
R_{d_1}	1.0	1.05	1.05	0.52
R_{d_2}	1.0	1.05	1.15	0.49

Table 2: Windkessel model with two distal resistances: σ_u^2 before and after considering mutual information between the two distal resistances

that R_{d_1} and R_{d_2} should have lower expected gain in information, but their mutual information should be high. This is exactly what is observed in the numerical results as shown in Figure 6. Compared to the previous example, where the individual gain in information was 1.2 nats, here R_{d_1} and R_{d_2} show an expected information gain of only 0.35 nats. On the other hand, the conditional mutual information between R_{d_1} and R_{d_2}

is now 0.22 nats, which is comparable to 0.35 nats in magnitude, implying that large correlations between these parameters exist. The results of EEV are shown in Table 2. It is observed that these parameters now depict an σ_e^2 of roughly the same magnitude as the prior variance. Furthermore, if one adds the conditional mutual information to the original individual gain (implying that one of the two parameters was known/fixed), the σ_e^2 drops to half of the original value – a significant increase in identifiability. This further confirms that correlation between R_{d_1} and R_{d_2} cannot be ignored.

Lastly, in Figure 4, the differences between the original estimator by Kraskov et. al. [25] and its modified version, where the distances in the largest space are used, are shown for the parameter R_p. It is observed that while the original estimator performs quite well, it shows an undesirable negative slope (decrease in entropy) in regions of time, $t \in [0.45, 0.5]$, where entropy remains constant. The modified estimator respects the non-decreasing nature of conditional entropy better than the original estimator.

7.2. SIR model.

In this section an application to a non-linear ODE model used in epidemiology (Susceptible-Infected-Recovered, SIR) is presented. This system has been studied in [6] by using a Gröbner basis approach.

The model is a non-linear compact SIR model of the form:

$$\dot{S} = \mu N - \mu S - \frac{\eta}{N} SI, \tag{67}$$

$$\dot{I} = -(\mu + \omega) I + \frac{\eta}{N} SI, \tag{68}$$

where μ, N, ω, η are scalar parameters, S and I are the number of susceptible and the infected individuals, respectively. The total number of imdividuals $N = S + I + R$ is a constant parameter for this model, where R is the number of recovered individuals. The state is $\mathbf{x} = (S, I)$ and the observation is a fraction of the infected individuals:

$$z = \kappa I, \tag{69}$$

where $\kappa \in [0, 1]$ is an unknown parameter. The initial conditions S_0, I_0 are unknown and can be considered as part of the parameters. Therefore, the parameters vector is: $\boldsymbol{\vartheta} = (\mu, N, \eta, \omega, \kappa, S_0, I_0)$.

The particularity of this model is that the parameter κ does not affect the dynamics, but only the observations. The level of noise models the number of false positive and negative arising in the diagnosis of the infected individuals that actually go to the medical doctor.

The choice for the prior is discussed hereafter. The total number of individuals has to be positive, so a lognormal prior is assumed of the form: $N = N_0 \cdot 2^{\vartheta_N}$, where $N_0 = 100$, ϑ_N is normally distributed with mean $\overline{\vartheta}_N = 0$ and variance $\sigma_N^2 = 0.05$. The initial conditions for the susceptible individuals (namely S_0) is chosen analogously. Let $S_0 = \gamma N$. A prior for the fraction γ is chosen to be $\gamma_0 \cdot 2^{\vartheta_\gamma}$, where and $\gamma_0 = 0.3$ and the exponent is a random variable normally distributed, with mean $\overline{\vartheta}_\gamma = 0$ and

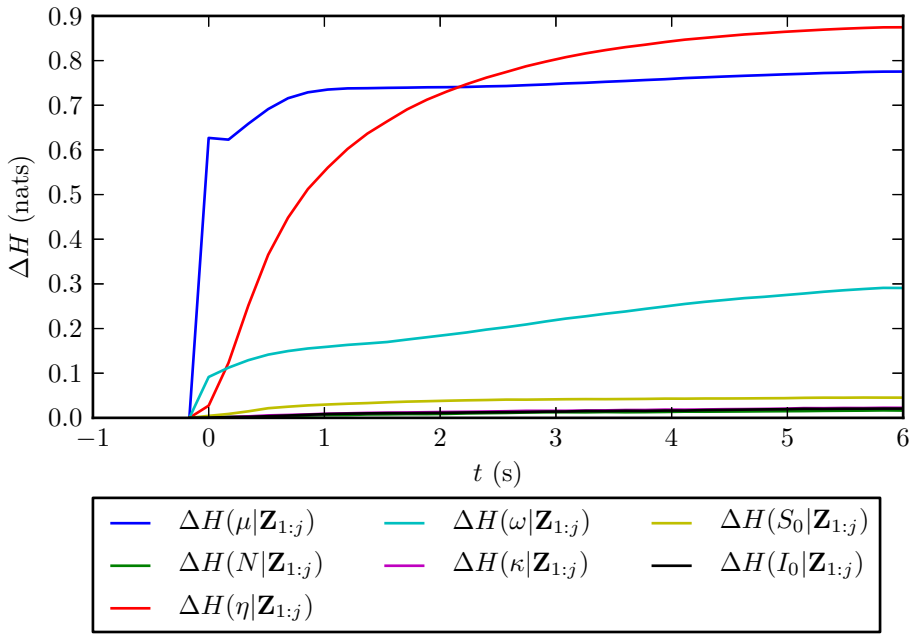


Figure 7: Expected information gains in (nats) for the SIR model (section 7.2) against time in (s) for all the parameters.

variance $\sigma_\gamma^2 = 0.05$. The prior for the infected individuals at initial time is chosen to be $I_0 = \alpha N$, where the prior for α is lognormal, $\alpha = \alpha_0 \cdot 2^{\vartheta_\alpha}$, with $\alpha_0 = 0.5$ and the exponent being normally distributed, with the same statistics as γ. The prior for κ is assumed to be gaussian, with mean $\overline{\kappa} = 0.4$, and variance $\sigma_k^2 = 0.1$. For the other variables, normal priors are assumed: for μ the mean value is $\overline{\mu} = 2.0$ and variance $\sigma_\mu^2 = 0.4$; for η: mean $\overline{\eta} = 1.75$, and variance $\sigma_\eta^2 = 0.25$; and for the parameter ω, mean $\overline{\omega} = 0.6$ and variance $\sigma_\omega^2 = 0.1$. The above choice of the priors is made to be reasonably consistent with respect to each other. For instance, the number of individuals in each of the susceptible, infected, and recovered, classes should be less than the total number of individuals.

A number of $N_{\text{ens}} = 10000$ samples is generated and 25 neighbours are used to approximate the entropies. The model is integrated in time by using a second order Crank-Nicolson scheme and 36 observations are retained, uniformly distributed between initial and final time ($T = 6$).

The expected information gain for the individual parameters are shown in Fig.7. From the plot it can be inferred that the expected information gains for the parameter N, S_0, I_0, and κ are negligible when compared to the other remaining parameters, hinting that they are not identifiable.

In Table 3 the results are shown in terms of EEVs, confirming that the parameters N, κ and the initial conditions for the population are not identifiable. The results obtained are in good agreement with those obtained in [6], showing that for this epidemiologic model, the parameters governing the disease dynamics (namely μ, η, ω)

ϑ	$\overline{\boldsymbol{\vartheta}}$ (prior mean)	σ^2_ϑ (prior variance)	σ^2_e	σ^2_u
μ	2.0	0.4	0.101	0.265
ϑ_N	0	0.05	8.50	120.07
η	1.75	0.25	0.054	0.213
ω	0.6	0.1	0.14	1.24
κ	0.4	0.1	0.801	20.82
ϑ_α	0	0.05	2.09	27.68
ϑ_γ	0	0.05	1.05	14.75

Table 3: Results for the SIR model (see Section 7.2). In the first column, the parameters, in the second and third column, the average and the variance of the prior, in the last column the variances of the EEVs.

can be identified, while the population characteristic sizes cannot.

Remark 7.1. It should be noted that the choice of priors can influence the outcome of identifiability. This is particularly true for this example where the parameter κ does not affect the dynamics but only the observations. From equation (69) it can be seen that if the prior on I_0 is strong, and the observations at the beginning of the experiment have low noise, then the posterior on κ is relatively precise irrespective of its prior. In such a case, which might not be physically reasonable in terms of the chosen priors, the parameter κ is identifiable.

7.3. Potential identification for harmonic waves.

An example where the system dynamics is governed by a PDE is presented in this section. The system is a 1 + 1-dimensional hyperbolic PDE defined for $(\xi, t) \in [0,1] \times [0,T]$, where $T = 4$ is the final time. The system describes harmonic waves in the presence of an unknown potential field. The system equation reads:

$$\partial_t^2 u = \partial_\xi^2 u + V(\xi) u, \tag{70}$$

where u is the displacement and homogeneous Dirichlet boundary conditions are imposed $u(0,t) = u(1,t) = 0$. The system has a deterministic initial condition $u(\xi, 0) = \sin(\pi\xi)$. The potential is parametrised by a harmonic expansion as follows:

$$V(\xi) = A \sum_j^{N_h} \hat{v}_j \sin(\pi j \xi), \tag{71}$$

where $N_h = 4$ is the maximum number of harmonics considered, $A = 20$ is a scaling coefficient.

The identifiability of different space wave numbers $\hat{v}_j$ is investigated. The measurements are the displacements u at several points in the domain. The direct simulation is performed by discretising in space by means of piecewise linear Finite Elements ($N_x = 256$ degrees of freedom), in time by an implicit second order Crank-Nicolson scheme, with $\Delta t = 8 \cdot 10^{-3}$. The observation consists in measuring the

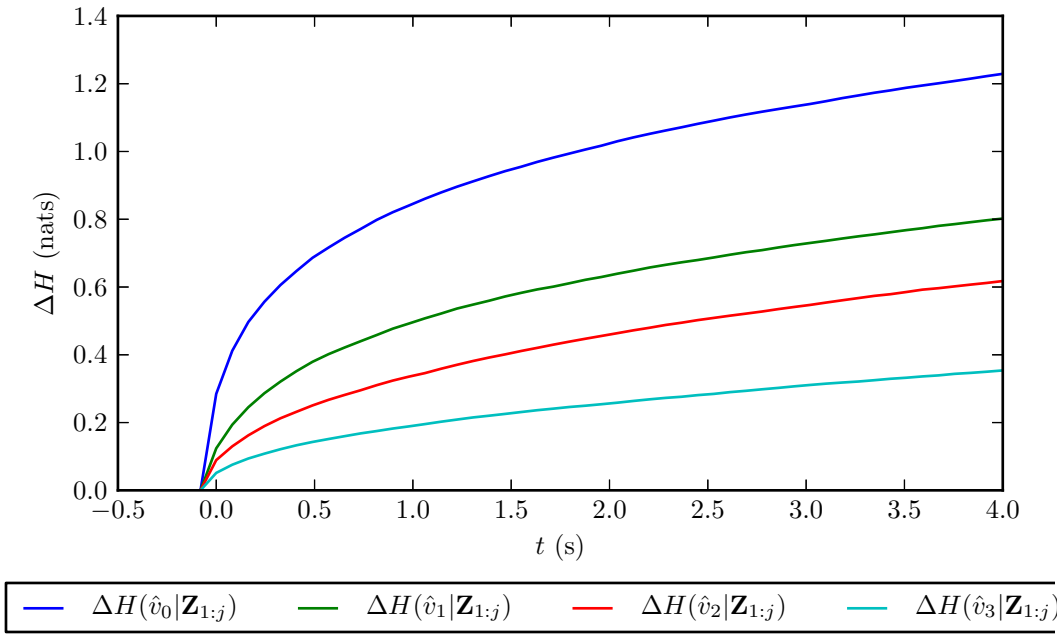


Figure 8: Information gains in (nats), for the 1D wave model (section 7.3) against time in (s) for all the harmonics of the potential.

position u in $N_z = 12$ points, uniformly distributed (excluding the extrema), with a time sampling frequency of $\nu = 12.5\ Hz$ (corresponding to 1 over 10 simulated time instants). By adopting this discretisation, the state $\mathbf{x} = (u(\xi_1), ..., u(\xi_{Nx}))$ is the value of the solution in the points of the uniform mesh, the observation $\mathbf{z} = (u(\xi_1^0), ..., u(\xi_{Nz}^0))$ is the value of the solution in the observation points, corrupted by a gaussian noise with zero mean and variance $\sigma_n^2 = 10^{-3}$. The parameter vector is $\boldsymbol{\vartheta} = (\hat{v}_1, \ldots, \hat{v}_{N_h})$. The joint prior for the parameters is assumed to be gaussian, with mean $\overline{\boldsymbol{\vartheta}}$ and covariance matrix $\mathbf{S}$, defined as:

$$\overline{\boldsymbol{\vartheta}} = (0, 0, 0, 0), \tag{72}$$

$$\mathbf{S} = \mathrm{diag}(1, 1, 1, 1). \tag{73}$$

In Fig.8 the expected information gains for the different harmonics are shown. For every parameter, most of the expected information gain occurs in the first part of the evolution, i.e. for $t < T_w/4$, where T_w is the time period of the wave. After $T_w/4$, the expected information gain stabilises and there is a continuous learning throughout the dynamical system evolution. This provides a useful information for the setting up of the measuring procedure. As it is expected, low space frequency harmonics are well identifiable, while high space frequencies are not. Indeed, given the observation process, there exists a cut-off on the space frequencies, which is determined by the number of space points monitored. Frequencies higher than $1/N_z$ are not identifiable by construction.

The results are summarised in Table 4. The EEV values clearly indicate that the identification is more and more difficult as the space frequency increases.

ϑ	$\overline{\boldsymbol{\vartheta}}$ (prior mean)	σ^2_ϑ (prior variance)	σ^2_e	σ^2_u
$\hat{v}_1$	0	1.0	0.094	0.094
$\hat{v}_2$	0	1.0	0.247	0.247
$\hat{v}_3$	0	1.0	0.417	0.417
$\hat{v}_4$	0	1.0	0.967	0.967

Table 4: Results for the Wave model (see Section 7.3). In the first column, the parameters, in the second and third column, the mean and the variance of the prior, in the last columns the mean value of the prior and the variances of the EEVs.

An analysis of the mutual information between the parameters is performed, showing that all the mutual informations are negligible relative to the values of the expected information gains for the single parameters. The interpretation is that the unidentifiability of high space frequencies does not depend on the correlation between the parameters, but is an intrinsic physical property of the system and the observation process adopted.

7.4. Advection-Difffusion equation in 2D.

The last example is a system governed by a $2+1$ dimensional PDE, that is aimed to mimic a source detection inverse problem for an advection-diffusion equation. A physical point domain is $(\boldsymbol{\xi}, t) = (\xi_1, \xi_2, t) \in \Omega = [0,1]^2 \times [0, 0.1]$, where ξ_1 and ξ_2 are the horizontal and vertical coordinates respectively. The system unknown is a passive scalar c that diffuses and is simultaneously advected by a velocity $\mathbf{v}$. The equation governing the dynamics reads:

$$\partial_t c + \mathbf{v} \cdot \nabla c = \mu \nabla^2 c, \tag{74}$$

$$c(\xi_1|_{(0,1)}, \xi_2, t) = 0 \ \text{ on } \partial\Omega_{lat}, \tag{75}$$

$$c(\xi_1, 0, t) = S(\xi_1) \ \text{ on } \partial\Omega_{bot}, \tag{76}$$

$$c(\boldsymbol{\xi}, 0) = 0. \tag{77}$$

where $S(\xi_1)$ is the source function, defined by:

$$S(\xi_1) = \begin{cases} s_1 \xi_1 (1/2 - \xi_1) & \xi_1 \leq 1/2 \\ s_2 (\xi_1 - 1/2)(1 - \xi_1) & 1/2 < \xi_1 \leq 1 \end{cases} \tag{78}$$

The intensities of the parabolic sources are positive scalar parameters (s_1, s_2). The diffusivity of the passive scalar c is denoted by μ. The velocity is mostly aligned with the ξ_2 axis, so that on the upper boundary characteristics are exiting and no boundary conditions need to be imposed. The probes are located at a constant value $\xi_2 = 0.6$, uniformly distributed along the horizontal coordinate. A number of $N_s = 20$ probes are considered. This setting is shown in Fig.9.

The equations are discretised by means of standard finite element methods, considering $N_{dof} = 10^4$ degrees of freedom in space. The equations are integrated in time by using a Crank-Nicolson scheme and taking 800 time steps.

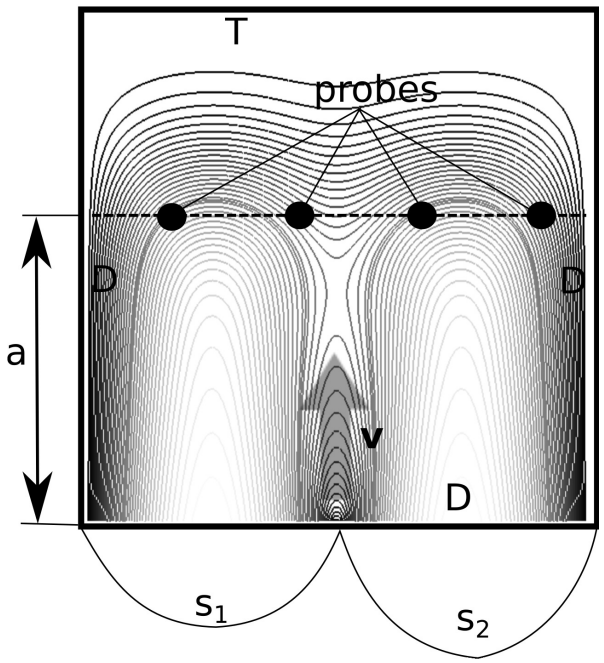


Figure 9: Setting for the PDE problem of advection-diffusion, see Sec.7.4. The domain is the unit square, the probes are located at $y = a = 0.6$. The capital letter D corresponds to Dirichlet boundary conditions while T are transparent boundary conditions. The velocity is mostly aligned with the ξ_2 axis. The function that mimics the source is characterised by two parameters $S_{1,2}$. Superimposed, the contours of the solution at $t = 5\ 10^{-2}$ normalised, 40 values between the maximum and the minimum. The parameters to obtain it are: $S_1, S_2 = 25$, $v_1 = 0$, $v_2 = 10$, $\mu = 0.1$.

The goal of the study is the identifiability of the parameters s_1, s_2, μ, v_1, and v_2 (where v_1 and v_2 are the components of the drift velocity) by measuring the concentration at the probe locations. Two different scenarios are investigated: a drift regime and a diffusive regime. It should be noted that these scenarios are implemented by simply changing the prior. The differences in terms of identifiability are assessed.

7.4.1. Drift regime: In this section, the drift dominated regime is investigated. Intuitively, owing to low diffusion, the information about the sources is transported downstream relatively unaffected to the probes. The prior is assumed as follows. The source intensities (s_1, s_2) and the diffusivity (μ) have a lognormal prior. For the sources, given $s_0 = 25$, the following parameterisation for priors for s_1 and s_2 are chosen: $s_1 = s_0 2^{\vartheta_{s_1}}$ and $s_2 = s_0 2^{\vartheta_{s_2}}$. Here, for the prior distribution, both ϑ_{s_1} and ϑ_{s_2} are normally distributed random variables with zero mean and variance 0.2. Similarly, $\mu = \mu_0 2^{\vartheta_\mu}$, where $\mu_0 = 0.1$, and for the prior distribution ϑ_μ is a normally distributed random variable with zero mean and variance $5{\times}10^{-2}$. The priors for v_1 and v_2 are assumed normal with means 0.0 and 10.0, and variances 0.05 and 1.0, respectively.

The results in terms of expected information gains are shown in Fig.10. The source

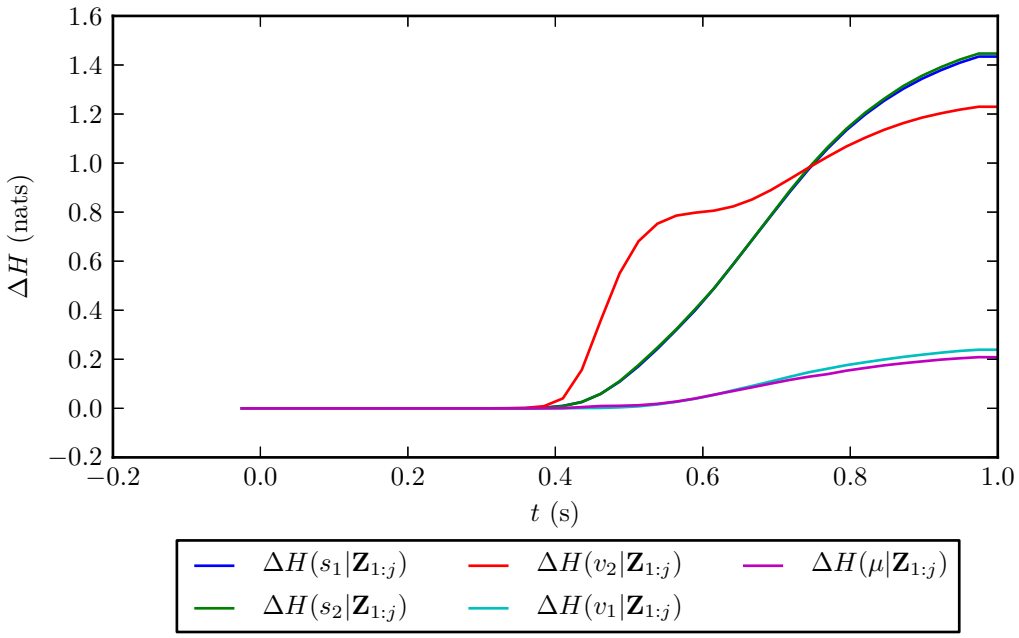


Figure 10: Expected information gains in (nats), for the advection-diffusion equation in the drift-dominated regime of motion (section 7.4.1) against time in (s) for all the parameters.

ϑ	$\overline{\boldsymbol{\vartheta}}$ (prior mean)	σ^2_ϑ (prior variance)	σ^2_e	σ^2_u
ϑ_{s_1}	0	0.2	0.01	0.059
ϑ_{s_2}	0	0.2	0.01	0.059
v_1	0.0	0.05	0.09	1.63
v_2	10.0	1	0.02	0.093
ϑ_μ	0	0.05	0.06	1.93

Table 5: Results for the advection-diffusion model in a drift regime (see Section 7.4.1). In the first column, the parameters, in the second and third column, the mean and the variance of the prior, in the last columns the mean value of the prior and the variances of the EEVs.

intensities are well identifiable. The less identifiable parameters are the diffusivity μ (which is quite small in the drift-dominated regime) and the horizontal velocity v_1.

It should be noted that in Figure 10 the expected gain in information concerning the vertical component of the velocity v_2 increases before that related to the source intensities s_1 and s_2. This is due to the fact that the front propagates from the bottom to the top, and, when it reaches the probes location, the change in the measurements can be first related to the advection speed and then to the intensity of the sources. The results are summarised in Table 5: the EEV analysis confirms that the source intensities and the drift-velocity v_2 are well identifiable.

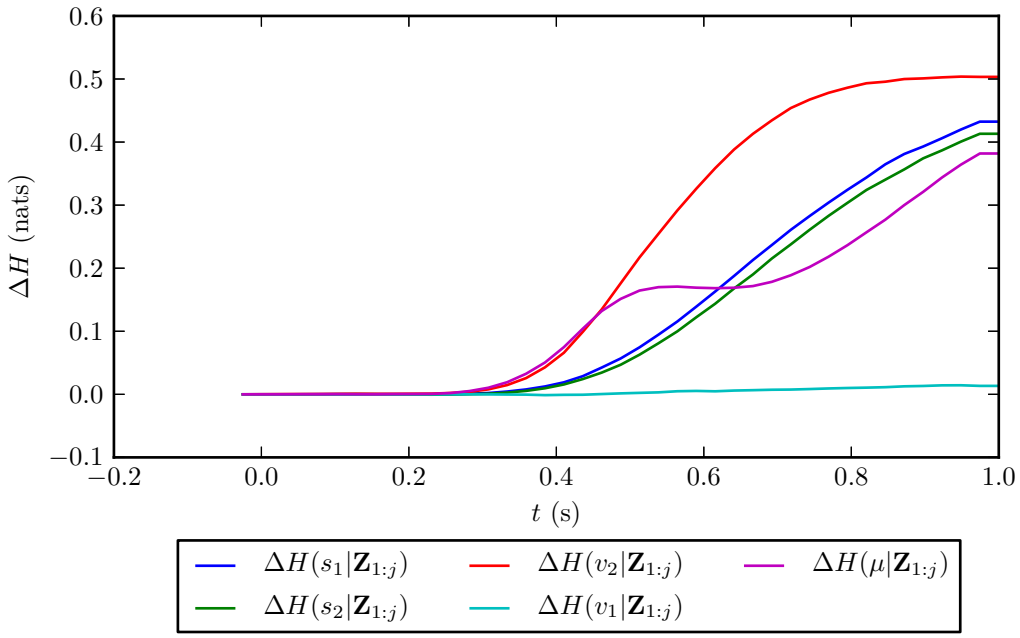


Figure 11: Expected information gains in (nats), for the advection-diffusion equation in the diffusion regime of motion (section 7.4.1) against time in (s) for all the parameters.

ϑ	$\overline{\boldsymbol{\vartheta}}$ (prior mean)	σ^2_ϑ (prior variance)	σ^2_e	σ^2_u
ϑ_{s_1}	0	0.2	0.14	0.72
ϑ_{s_2}	0	0.2	0.15	0.78
v_1	0.0	0.05	1.74	35.7
v_2	7.5	1	0.58	0.59
ϑ_μ	0	0.05	0.01	0.86

Table 6: Results for the advection-diffusion model in a diffusive regime (see Section 7.4.2). In the first column, the parameters, in the second and third column, the mean and the variance of the prior, in the last columns the mean value of the prior and the variances of the EEVs.

7.4.2. Diffusive regime: In this section a diffusive regime, where the passive scalar dynamics is dominated by diffusion, is considered. All the parameters are kept constant as in the previous section except those for diffusivity and drift velocity. For μ the parameterisation $\mu = \mu_0 2^{\vartheta_\mu}$ is considered, where $\mu_0 = 1.5$ and ϑ_μ is a normally distributed random variable with zero mean and variance 10^{-2}. For v_2 a normal distribution with mean 7.5 and variance 1.0 is considered.

The expected information gains for these settings are shown in Fig.11. The sources are still individually identifiable. This is due to the fact that the drift velocity is still high. However, the gain in information about the source intensities is much less when compared to the drift-dominated case (0.4 nats over 1.4). These results are confirmed by the values reported in Table 6. The values of the EEVs are about 3 times larger.

Even if in both cases the source intensities are mathematically identifiable in a noiseless setting, there is a difference from a pragmatical standpoint. In the presence of noise, when different regimes of motion are considered, the information learned about the parameters can vary significantly. This is an important contribution of this kind of analysis.

Remark 7.2 (Limitations of the $k-$nearest neighbour mutual information estimator)**.** While the estimator for mutual information and conditional mutual information works well for the above presented examples, it suffers from a drawback. In particular, in equation (26), if the joint probability distribution of $\mathbf{\Theta}$ and $\mathbf{Z}_{1:j}$ lie on a lower-dimensional manifold, then the estimator yields erroneous results due to the incorrect assumption of constant uniform density in the ϵ-balls; see [25] for this assumption in the formulation of the estimator and [36] for a discussion on this drawback. To ensure that such a drawback does not effect the above presented results, the results of the k-nearest neighbour based estimator are validated against those obtained by other methods, such as kernel density estimation.

8. Conclusion and Perspectives.

In this work a semi-empirical method to assess practical identifiability of parametric dynamical systems is proposed . It is set in a Bayesian framework and consists of analysing several information-theoretic measures. In particular, decrease in uncertainty of the system parameters due to the availability of noisy system measurements is computed via Shannon entropy. This decrease is viewed as the expected information gain and is related to identifiability of the parameters. Furthermore, the concept of a hypothetical measurement device that measures the parameters directly is proposed to facilitate interpretation of uncertainty reduction. In order to determine which subsets of parameters can only be identified collectively (i.e. individual parameters in this subset are unidentifiable), conditional mutual information is employed.

Four advantages of the proposed method are particularly relevant: first, that the method can take the measurement-noise of any structure into account; second, that it can be applied to generic dynamical systems including those governed by PDEs; third, that it is non-intrusive and does not require any modification of the existing numerical codes to solved the dynamical systems; and finally, that it has the ability to differentiate between parameter identifiability in different regions of the parametric space. The application of this approach is presented in a range of dynamical systems (including non-linear ODEs and PDEs) and, where possible, results are validated against those independently published in the literature.

There are two main drawbacks of the proposed method. First, that it is computationally intensive; however, this drawback is partly mitigated by the easily parallelisable nature of the method. Second, that the employed mutual information estimator presents problems if the joint distribution of the parameters and the observables lies in a low-dimensional manifold. To avoid the latter problem, future work

includes assessment and development of non-parametric entropy estimation methods in manifolds (see for example [37]).

References

[1] Bellman R and Astrom K 1970 *Mathematical Biosciences* **7** 329–339
[2] Cobelli C and Di Stefano J 1980 *American Journal of Physiology* **239(1)** n/a–n/a
[3] Miao H, Xia X, Perelson A S and Wu H 2011 *SIAM review* **53** 3–39
[4] Pohjanpalo H 1978 *Mathematical Biosciences* **41** 21–33
[5] Nemcova J 2010 *Mathematical Biosciences* **223** 83–96
[6] Meshkat N, Eisemberg M and DiStefano III J J 2009 *Math. Biosci.* **222** 61–72
[7] Denis-Vidal L, Joly-Blanchard G and Noiret C 2001 *Mathematics and computers in simulation* **57** 35–44
[8] Raue A and al 2009 *Bioinformatics* **25** 1923–1929
[9] Kaipio J and Somersalo E 2004 *Statistical and Computational inverse problems* (Springer Verlag, New York)
[10] Oden T, Prudencio E and Hawkins-Daarud A 2013 *Mathematical Models and Methods in Applied Sciences* n/a–n/a
[11] Huan X and Marzouk Y 2013 *Journal of Computational Physics* **231** 288–317
[12] Liepe J, Filippi S, Komorowski M and Stumpf M 2013 *Plos Computational Biology* **9** n/a–n/a
[13] Eisenberg M and Hayashi M 2014 *Mathematical Biosciences* **256** 116–126
[14] Nobile F, Babuska I and Tempone R 2010 *Siam Review* **52** 317–355
[15] Fishman G 1996 *Monte Carlo: concepts, algorithms and applications* (Springer Verlag, New York)
[16] Xiu D and Karniadakis G E 2002 *Siam Journal of Scientific Computing* **24** 619–644
[17] Shannon C 1948 *The Bell System Technical Journal* **27** 379–423
[18] Jaynes E T 1963 Information Theory and Statistical Mechanics (Notes by the lecturer) *Statistical Physics 3 (1962 Brandeis Lectures)*
[19] Kullback S and Leibler R A 1951 *The Annals of Mathematical Statistics* **22** 79–86
[20] Kullback S 1997 *Information theory and statistics* (Courier Dover Publications)
[21] Ghosh S, Burnham K P, Laubscher N F, Dallal G E, Wilkinson L, Morrison D F, Loyer M W, Eisenberg B, Kullback S, Jolliffe I T and Simonoff J S 1987 *The American Statistician* **41** pp. 338–341
[22] Hobson A and Cheng B K 1973 *Journal of Statistical Physics* **7** 301–310
[23] Hobson A 1969 *Journal of Statistical Physics* **1** 383–391
[24] Cover T M and Thomas J A 2012 *Elements of information theory* (John Wiley & Sons)
[25] Kraskov A, Stögbauer H and Grassberger P 2004 *Phys. Rev. E* **69**(6) 066138
[26] Jensen J 1906 *Acta Mathematica* **30** 175–193
[27] Beirlant J, Dudewicz E J, Györfi L and Van der Meulen E C 1997 *International Journal of Mathematical and Statistical Sciences* **6** 17–39
[28] Steuer R, Kurths J, Daub C O, Weise J and Selbig J 2002 *Bioinformatics* **18** S231–S240
[29] Moon Y I, Rajagopalan B and Lall U 1995 *Physical Review E* **52** 2318
[30] Kozachenko L and Leonenko N 1987 *Problems Inform. Transmission* **23** 9–16
[31] Singh H, Misra N, Hnizdo V, Fedorowicz A and Demchuk E 2003 *American journal of mathematical and management sciences* **23** 301–321
[32] Westerhof N, Lankhaar J W and Westerhof B 2009 *Medical & Biological Engineering & Computing* **47** 131–141
[33] Vignon-Clementel I E, Alberto Figueroa C, Jansen K E and Taylor C a 2006 *Computer Methods in Applied Mechanics and Engineering* **195** 3776–3796
[34] Thomaseth K and Cobelli C 1999 *Annals of Biomedical Engineering* **27** 607–616
[35] Pant S, Fabrges B, Gerbeau J F and Vignon-Clementel I E 2014 *International Journal for Numerical Methods in Biomedical Engineering* **30** 1614–1648

[36] Greg Ver S 2014 *Online documentation* Available at: `http://www.isi.edu/~gregv/npeet_doc.pdf`, last accessed on 12 Dec 2014
[37] Costa J A and Hero A O 2004 *Signal Processing, IEEE Transactions on* **52** 2210–2221

Appendix A. Different measures of uncertainty lead to the same expected information gain

This appendix presents details of the argument that irrespective of which measure of inherent uncertainty is used, the expected gain in information when the probability distribution changes from $p_X(x)$ tp $p_{X|Y}(x|y)$ remains the same.

Appendix A.1. Shannon uncertainty

If equation (2) is seen as an inherent measure of uncertainty then the gain in information $G_{X|Y=y}(y)$ from equation (8) is

$$G_{X|Y=y}(y) = \int_{\mathcal{X}} p_X(x)\ \log\left(\frac{1}{p_X(x)}\right)\ dx\ -\ \int_{\mathcal{X}} p_{X|Y}(x|y)\ \log\left(\frac{1}{p_{X|Y}(x|y)}\right)\ dx. \quad \text{(A.1)}$$

Consequently the expected gain in information, $I^S_{X|Y}$, from equation (9) is

$$I^S_{X|Y} = \int_{\mathcal{X}} p_X(x)\ \log\left(\frac{1}{p_X(x)}\right)\ dx - \int_{\mathcal{Y}}\int_{\mathcal{X}} p_{X|Y}(x|y)\ \log\left(\frac{1}{p_{X|Y}(x|y)}\right) p_Y(y)\, dx\, dy. \quad \text{(A.2)}$$

$$I^S_{X|Y} = \int_{\mathcal{X}} p_X(x)\ \log\left(\frac{1}{p_X(x)}\right)\ dx\ -\ \int_{\mathcal{Y}}\int_{\mathcal{X}} p_{X,Y}(x,y)\ \log\left(\frac{1}{p_{X|Y}(x|y)}\right)\ dx\ dy. \quad \text{(A.3)}$$

where $p_{X|Y}(x,y)$ represents the joint distribution of X and Y.

Appendix A.2. Jaynes' uncertainty

If equation (3) is seen as an inherent measure of uncertainty then the gain in information $G_{X|Y=y}(y)$ from equation (8) is

$$G_{X|Y=y}(y) = \int_{\mathcal{X}} p_X(x)\ \log\left(\frac{m(x)}{p_X(x)}\right)\ dx\ -\ \int_{\mathcal{X}} p_{X|Y}(x|y)\ \log\left(\frac{m(x)}{p_{X|Y}(x|y)}\right)\ dx. \quad \text{(A.4)}$$

Since $m(x)$ is an invariant measure of the random space and, it remains constant. Consequently, $I^J_{X|Y}$, can be written from equation (9) as

$$I^J_{X|Y} = \int_{\mathcal{X}} p_X(x)\ \log\left(\frac{m(x)}{p_X(x)}\right)\ dx - \int_{\mathcal{Y}}\int_{\mathcal{X}} p_{X|Y}(x|y)\ \log\left(\frac{m(x)}{p_{X|Y}(x|y)}\right) p_Y(y)\, dx\, dy \quad \text{(A.5)}$$

$$I^J_{X|Y} = I^s_{X|Y}\ +\ \int_{\mathcal{X}} p_X(x)\ \log\left(m(x)\right)\ dx\ -\ \int_{\mathcal{Y}}\int_{\mathcal{X}} p_{X,Y}(x,y)\ \log\left(m(x)\right)\ dx\ dy. \quad \text{(A.6)}$$

Since $\int_{\mathcal{Y}} p_{X,Y}(x,y)\ dy = p_X(x)$, the last two terms in the above equation cancel and hence

$$I^J_{X|Y} = I^S_{X|Y} \quad \text{(A.7)}$$

Appendix A.3. Kullback-Leibler distance

According to the interpretation of the Kullback-Leibler distance that $D(p_X(x)||q_X(x))$ represents the gain in information when the probability distribution changes from $q_X(x)$ to $p_X(x)$. Hence, $G_{X|Y=y}(y)$ from equation (4) is

$$G_{X|Y=y}(y) = D(p_{X|Y}(x|y)||p_X(x)) \tag{A.8}$$

$$= \int_{\mathcal{X}} p_{X|Y}(x|y) \ \log\left(\frac{p_{X|Y}(x|y)}{p_X(x)}\right) \ dx. \tag{A.9}$$

Consequently, $I^K_{X|Y}$ can be written from equation (9) as

$$I^K_{X|Y} = \int_{\mathcal{Y}}\int_{\mathcal{X}} p_{X|Y}(x|y) \ \log\left(\frac{p_{X|Y}(x|y)}{p_X(x)}\right) p_Y(y) \ dx \ dy \tag{A.10}$$

$$= \int_{\mathcal{Y}}\int_{\mathcal{X}} p_{X,Y}(x,y) \ \log\left(\frac{1}{p_X(x)}\right) \ dx \ dy \ - \ \int_{\mathcal{Y}}\int_{\mathcal{X}} p_{X,Y}(x,y) \ \log\left(\frac{1}{p_{X|Y}(x|y)}\right) \ dx \ dy \tag{A.11}$$

$$= \int_{\mathcal{X}} p_X(y) \ \log\left(\frac{1}{p_X(x)}\right) \ dx \ dy \ - \ \int_{\mathcal{Y}}\int_{\mathcal{X}} p_{X,Y}(x,y) \ \log\left(\frac{1}{p_{X|Y}(x|y)}\right) \ dx \ dy \tag{A.12}$$

$$= I^S_{X|Y} \tag{A.13}$$

Appendix A.4. Hobson uncertainty

If equation (5) is seen as an inherent measure of uncertainty then the gain in information $G_{X|Y=y}(y)$ from equation (8) is

$$G_{X|Y=y}(y) = \left(D(p^m_X(x)||p^0_X(x)) - D(p_X(x)||p^0_X(x))\right) \tag{A.14}$$

$$- \left(D(p^m_X(x)||p^0_X(x)) - D(p_{X|Y}(x|y)||p^0_X(x))\right)$$

$$= D(p_{X|Y}(x|y)||p^0_X(x)) - D(p_X(x)||p^0_X(x)) \tag{A.15}$$

$$= \int_{\mathcal{X}} p_{X|Y}(x|y) \ \log\left(\frac{p_{X|Y}(x|y)}{p^0_X(x)}\right) \ dx \ - \ \int_{\mathcal{X}} p_X(x) \ \log\left(\frac{p_X(x)}{p^0_X(x)}\right) \ dx \tag{A.16}$$

Consequently, $I^H_{X|Y}$ can be written from equation (9) as

$$I^H_{X|Y} = \int_{\mathcal{Y}}\int_{\mathcal{X}} p_{X|Y}(x|y) \ \log\left(\frac{p_{X|Y}(x|y)}{p^0_X(x)}\right) p_Y(y) \, dx \, dy - \int_{\mathcal{X}} p_X(x) \ \log\left(\frac{p_X(x)}{p^0_X(x)}\right) \ dx \tag{A.17}$$

$$I^h_{X|Y} = \int_{\mathcal{Y}}\int_{\mathcal{X}} p_{X,Y}(x,y) \ \log\left(\frac{p_{X|Y}(x|y)}{p^0_X(x)}\right) \ dx \ dy \ - \ \int_{\mathcal{X}} p_X(x) \ \log\left(\frac{p_X(x)}{p^0_X(x)}\right) \ dx \tag{A.18}$$

$$= \int_{\mathcal{Y}}\int_{\mathcal{X}} p_{X,Y}(x,y) \ \log\left(p_{X|Y}(x|y)\right) \ dx \ dy \ - \ \int_{\mathcal{X}} p_X(x) \ \log\left(p_X(x)\right) \ dx$$

$$- \int_{\mathcal{Y}}\int_{\mathcal{X}} p_{X,Y}(x,y) \ \log\left(p^0_X(x)\right) \ dx \ dy \ + \ \int_{\mathcal{X}} p_X(x) \ \log\left(p^0_X(x)\right) \ dx \tag{A.19}$$

Since $\int_{\mathcal{Y}} p_{X,Y}(x,y)\ dy = p_X(x)$, the last two terms in the above equation cancel and hence

$$I^H_{X|Y} = \int_{\mathcal{X}} p_X(x)\ \log\left(\frac{1}{p_X(x)}\right)\ dx\ -\ \int_{\mathcal{Y}}\int_{\mathcal{X}} p_{X,Y}(x,y)\ \log\left(\frac{1}{p_{X|Y}(x|y)}\right)\ dx\ dy \quad \text{(A.20)}$$

$$= I^S_{X|Y} \quad \text{(A.21)}$$